# Prediction, Control, and the Regulating Mind

## A Proposed Model of Regulation, from Outside the Field

Dr. Greg Moody

August 17, 2026

## Contents

## 1. What I'm claiming, and what I'm not

Before I was a licensed psychotherapist and ran businesses and did martial arts and blah blah blah… I was an aerospace engineer. Guidance, navigation, control. Feedback loops with an estimate on one end and a correction on the other, and a short list of ways they fail.

I'm not a research psychologist. I don't run a lab, and I haven't spent thirty years inside this literature. A document from me proposing a model of psychology reads like somebody wandering in from another building to announce he's figured the place out. Fair enough. That isn't what this is.

The claim is smaller than the title makes it sound.

Two bodies of work already exist. One says the mind is a control system defending targets. The other says the mind is a prediction machine guessing at the world. Each one has a hole exactly where the other has something solid, and as far as I can find nobody has bolted them together to see what falls out of the join.

When you do, you get a handful of parameters that come apart cleanly. Those parameters predict different things about who gets better with which kind of help. That's the argument.

Why would somebody outside the field notice a join? Not because he's smarter. Because from inside either one, the other is somebody else's vocabulary. I spent years in one and then years in the other, so the fit was the first thing I saw instead of the last. Most cross-field ideas are wrong. This one might be too, and I can't tell that from where I'm standing.

Which is why Section 9 is in here. It lists, up front, the findings that would show this is wrong. If you read one part skeptically, read that one. I'd rather somebody knocked this down than nodded at it, and the people who could knock it down are the ones who've spent a career in one half of it.

> *A note on scope. I worked this out on a hard case first, autism and repetitive behavior, and there's a companion pair of documents covering that in full, including a long look at how thin the evidence there actually is. None of it is repeated here.*

# 2. The Mind as a Control System

## 2.1 The thermostat

Before I was a licensed psychotherapist, I was an aerospace engineer. Sensors, set points, feedback, gain, damping, and the ugly things a loop does when you get those wrong: that was my job description for years before it was ever a way of thinking about the person sitting across from me. Nothing in this section is a metaphor I went looking for. This is my first profession, explained to people in my second one.

Start with the device on the wall of almost every building you've ever worked in. It explains more about human behavior than most of what you were taught in graduate school, and it's worth taking apart slowly, because every idea that follows is already inside it.

A thermostat has a small piece of metal, or a small electronic component, that responds to the temperature of the air around it. That component is the **sensor**. It doesn't know the temperature of the room. It knows its own state, and its own state happens to vary with the temperature of the air touching it. Whatever number the thermostat displays belongs to the thermostat, not to the room. Call that reading the **perception**.

The distinction sounds pedantic until you mount the thing above a radiator. Now it reports a warm house on a freezing day and lets the pipes burst while insisting everything is fine. The device isn't lying. It's controlling exactly what it can measure, and what it can measure was never the same thing as the state of the world.

Somebody has set the thermostat to sixty-eight. That number is the **set point**. It makes no prediction about what the temperature will be and no claim about what the temperature is. It specifies what the temperature OUGHT to be, held internally, and defended.

Inside the thermostat is a piece of circuitry whose only job is to subtract one number from the other. That's the **comparator**, and the number it produces is the **error signal**. Set point sixty-eight, perception sixty-four, error four degrees, direction too cold. When the perception equals the set point the error is zero and the system does nothing at all. Zero error means the system already has what it's defending. That's what rest looks like in a control loop.

That single relation is the whole of it, and it's worth setting out on its own:

> **Error = set point − perception.** The system is driven not by the world, and not by its goal, but by the gap between the two.

The error signal goes to the **control element**: the furnace, the compressor, the valve. The control element acts on the world in whatever direction shrinks the error. Heat comes on, air warms, the sensor reads higher, the perception rises toward the set point, the error falls toward zero, the furnace shuts off. The loop closes on itself. Nothing in it required the thermostat to know anything about weather, insulation, or the family's habit of leaving the back door open.

Which brings us to **disturbances**. The back door. The January wind. The oven at four hundred degrees. The twelve people who came for dinner. The west-facing window at four in the afternoon. None of these is represented anywhere in the thermostat. It has no model of them and needs none. It notices only that its perception has moved, and it pushes back.

That property is why control systems are worth a clinician's attention. A control system produces stable outcomes in an unstable world while knowing nothing about the world. Its behavior is wildly variable, furnace running three minutes today and ninety minutes tomorrow, precisely BECAUSE its perception is being held constant. Variable action in the service of a constant perception is the signature of control, and to an observer who doesn't know what's being controlled, it looks like noise.

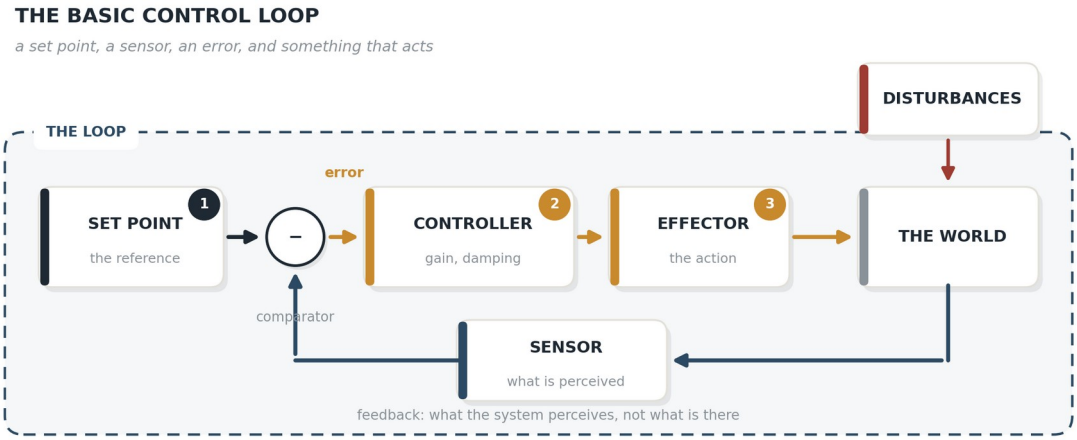


*Figure 2.1. The basic control loop.*

The vocabulary, in one place:

**Sensor.** The part that responds to the world. It reports its own state and nothing else. **Perception.** The number the sensor produces. The system's only access to reality. **Set point.** The value the system is trying to make its perception equal. **Comparator.** The circuit that subtracts one from the other. **Error.** What comes out of that subtraction. It has a size and a direction. **Control element.** The furnace. The part that acts on the world. **Disturbance.** Everything pushing the variable around that the system knows nothing about.

Three further properties turn that diagram into something that can go wrong in interesting ways.

**Gain** is how hard the system corrects for a given amount of error. A high-gain thermostat answers a one-degree deficit by running the furnace flat out. A low-gain thermostat answers the same deficit with a trickle of warm air. Gain has nothing to do with motivation and nothing to do with caring. It's a multiplier. Roughly: **output = gain × error.** High gain

buys tight control and fast recovery. It also buys overshoot, because by the time the room reaches sixty-eight the ducts are full of hot air that hasn't arrived yet, and the room sails past to seventy-one. Then the system corrects downward, overshoots again, and you get a house that's never comfortable and never still.

**Damping** is how much the system waits, how much it discounts the error it can see in favor of the correction already in flight. An undamped high-gain loop oscillates. A well-damped loop approaches the set point and settles. An overdamped loop is sluggish and never quite arrives, drifting a degree low forever because it won't push hard enough to close the last of the gap. Every clinician has met all three houses.

**Deadband**, sometimes called sensitivity or tolerance, is how much deviation the system will accept before it acts at all. A thermostat with a two-degree deadband set to sixty-eight does nothing between sixty-seven and sixty-nine. Widen the deadband and the system becomes tolerant, cheap to run, and imprecise. Narrow it to nothing and the system becomes exquisitely accurate and starts short-cycling itself to death, kicking on and off every ninety seconds, wearing out its own machinery.

Notice what's just been described without one psychological word: a system that's stable but not rigid, that can be too reactive or too slow, that can chase its own tail, that can be exhausted by its own sensitivity, and that can be catastrophically wrong about the world while performing perfectly by its own internal standard. Your Instant Pot has most of that architecture, and your Instant Pot has no inner life whatsoever.

Hold onto that last property…

A sensor mounted above a radiator gives you a working thermostat fed bad input. No amount of adjustment to the furnace will fix it.

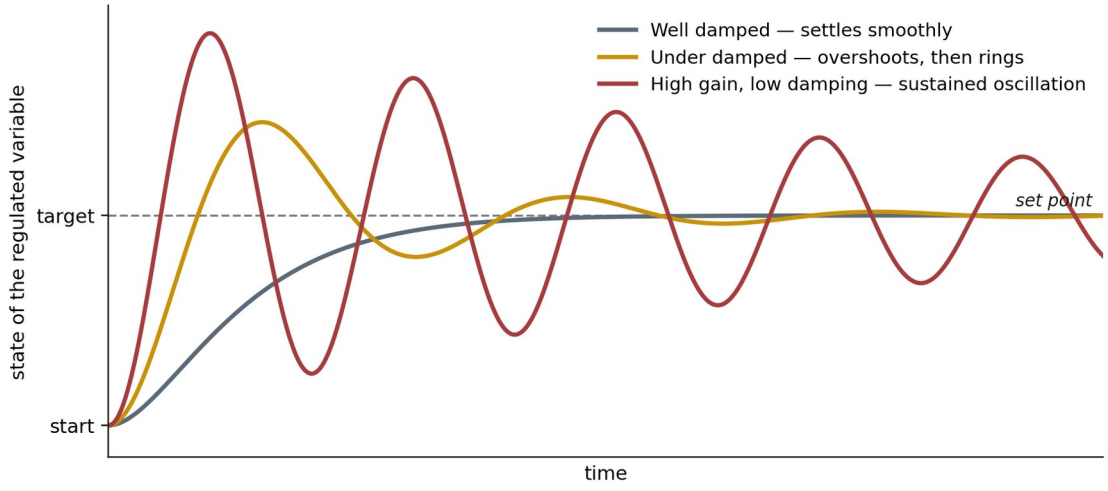


*Figure 2.2. Gain and damping: how a control loop settles, overshoots, or oscillates.*

## 2.2 Biological control

Walter Cannon gave this arrangement its physiological name almost a century ago. Writing in *Physiological Reviews*, he described the "organization for physiological homeostasis", the coordinated set of processes that hold the internal conditions of a mammal within narrow limits despite an environment that does nothing to help (Cannon, 1929). Body temperature, blood pH, calcium, glucose, oxygen saturation: each one is sensed, each one is compared against something like a reference, and each one is defended.

Take thirst, because it's the cleanest worked example in the body and because it will do a lot of work later.

The regulated quantity is the concentration of solutes in the blood, **plasma osmolality**. The sensors are osmoreceptive neurons in circumventricular organs of the forebrain, structures that sit outside the blood–brain barrier precisely so they can taste the blood directly (Zimmerman, Leib, & Knight, 2017). The set point is roughly 280 to 295 milliosmoles per kilogram, and the tolerance around it is startlingly tight. A rise of one or two percent is enough to produce a measurable response.

When osmolality drifts above that band, the comparison produces an error signal, and that error signal is **thirst**. Thirst carries no description of the blood and no information about water. What it carries is a magnitude with a direction and an unpleasantness, and it grows as the discrepancy grows. The output is behavior: get up, find water, drink it.

Then something worth pausing on happens. Thirst switches off within a minute or two of drinking, long before a drop of that water has been absorbed and long before plasma osmolality has moved at all. The oropharyngeal and gastric signals of swallowing feed forward into the same circuit and shut the error down in anticipation (Zimmerman et al., 2017; Gizowski & Bourque, 2018). Feedback alone is too slow here, so the body bolts feed-forward correction onto the comparator. Any engineer would recognize the fix, and it's the same reason a well-designed thermostat shuts the furnace off slightly early. We'll meet this idea again, under a different name, in Section 3.

Here's the move that matters, and it isn't a small one. The pancreas regulates blood glucose by secreting hormones. The kidney regulates sodium by adjusting excretion. These organs correct errors *inside the body*, using internal machinery, and you feel nothing. The brain regulates too, but the brain's control element is largely **the organism's behavior in the world**. When plasma osmolality rises, no internal mechanism can manufacture water. The only available correction is to move a body across a room, operate a tap, and swallow. The error has to be made available to whatever system chooses actions, and it has to be made unpleasant enough to win against everything else that system might do instead.

That's the design constraint. A biological control system whose only effective output is behavior can't afford a silent error signal.

## 2.3 The claim that emotions are error signals

In early 2025, the anonymous authors writing under the name Slime Mold Time Mold published a serial on their blog called *The Mind in the Wheel*. It ran to fourteen installments, a prologue, twelve numbered parts, and an interlude, between February and May of that year (Slime Mold Time Mold, 2025a–2025e). It's self-published, it hasn't been peer reviewed, and the authors describe it, without much modesty, as psychology's first paradigm. Take that framing for what it is. The core proposal still deserves a clinician's attention, and it's stated with unusual clarity.

Their claim, in their own words: *"An emotion is the error signal in a behavioral biological control system"* (Slime Mold Time Mold, 2025b).

The qualifier carries the argument. Serum copper is regulated. It's sensed, compared, and defended, and there's no such thing as feeling copper-deficient, not as an experience. The reason, on this account, is that copper regulation requires no behavior. It's handled internally, by processes the person never becomes aware of and that may not involve the brain at all. Thirst, hunger, cold, pain, the need to urinate, air hunger, fear, shame, disgust, loneliness: these are the errors of control systems whose corrections have to be carried out by the whole animal, and so they're the errors you can feel. Feeling, on this proposal, is what an error signal is like when it has to recruit behavior.

The vocabulary is deliberately mechanical, and it's worth learning as a set:

**Governor.** One complete control system. Sensor, comparator, set point, error, output. A person runs many at once, each defending its own variable toward its own target. **Selector.** The arbitrator. A body can only do one thing at a time, so something has to decide which error gets the body. **Vote.** What each governor casts for the behaviors that would reduce its error. The size of the vote scales with the size of the error. **Weight.** How loud a given governor gets to be next to its neighbors. Constitutional, not situational. **Gate.** A threshold that discards votes below it, which is why faint background hunger produces no behavior at all until it crosses a line and then abruptly does. **Knockout.** Their term for the natural experiment in which a governor is absent or silenced in a particular individual.

The knockout case they lean on is the patient with bilateral amygdala damage who reports essentially no fear yet still panics under carbon dioxide inhalation, which they read as evidence that fear and air hunger are separate governors rather than two shades of one thing (Slime Mold Time Mold, 2025e).

The architecture implies a taxonomy of failure, and this is where the framework starts to earn its keep clinically. A control loop has four places to break, and each break has a different signature.

DON'T ask whether the person in front of you is regulating. They are. Regulation is what a nervous system does.

ASK which of the four parts is producing the failure.

A **sensor fault** means the measurement is wrong. The variable is fine and the reading isn't. The system corrects vigorously and correctly toward a state of the world that doesn't exist.

A **comparator fault** means the subtraction is wrong. Sensor and set point are both accurate, and the error that comes out is too large, too small, or absent. Errors turned down by half give you a person who has to be twice as far out of alignment before they act. Errors turned down to near zero give you a person who does almost nothing.

A **set-point fault** means the target itself has moved. Nothing is broken in the machinery. The machinery is defending the wrong value, correctly and tirelessly.

An **output fault** means the error is generated accurately and the correction can't be executed. The furnace is disconnected. The person knows exactly what they need and can't make the behavior happen.

Four faults, one presentation. From outside, all four look like a system that isn't doing its job.

I have sat with clients in all four categories, and for years I read the flat ones as a motivation problem. That was a sensor fault of my own. The four faults call for four different interventions, and exactly ONE of them has anything to do with motivation.

Their application to depression follows from this. In Part V they work through seven named classes of malfunction that would all be diagnosed as depression: too much success (a life so well-regulated that errors are never large, so corrections are never large, so the person is never happy); happiness not generated (corrections occur normally but the mechanism that registers them fails, producing the patient who functions and describes himself as dead inside); errors not generated; errors generated too much; errors reduced by other routes; failures in voting and gating; and failures in what they call pricing, how the system weighs the cost of an action against the error it would relieve (Slime Mold Time Mold, 2025c). Mechanically these are unrelated conditions with unrelated treatments. Symptomatically they converge.

The observation they build on is a real one, and clinicians know it well. Most illnesses produce a symptom or fail to. Depression produces a symptom *and its opposite*. Insomnia and hypersomnia. Weight gain and weight loss. Eating too much and eating too little. Agitation and retardation. Overwhelming pain and total numbness. Their explanation is that these pairs are governed by two opposing systems rather than one: a governor that makes sure you sleep enough and a separate governor that makes sure you wake, a governor for eating enough and a governor against eating too much. As Powers argued long before them, a neural comparator can only signal error in one direction, so bidirectional control requires two loops (Slime Mold Time Mold, 2025b). A fault in the shared machinery can land on either side of a pair. Hit one and you get insomnia, hit the other and you get hypersomnia, and the underlying lesion is the same kind of lesion.

Their summary judgment on symptom-cluster diagnosis is the line most likely to stay with you. Imagine bringing your car to a mechanic who tells you it has a case of "broken-downness." You'd know immediately that he had no idea what he was talking about.

Broken-downness is an abstraction. It doesn't refer to anything and it won't help you fix a car. A competent mechanic says: your spark plugs are shot, so they can't drive the pistons. Then they observe that when a person feels sad all the time, we tell them they have depression, a category that carries no account of what parts are involved or what rules they follow (Slime Mold Time Mold, 2025a).

That's uncomfortable, and it isn't entirely fair to a diagnostic system built to do a different job. It also isn't entirely wrong.

## 2.3a Then what is happiness?

It's the first question anybody asks of that claim, and it's a good one. Hunger, fear and shame all behave like error signals. Joy doesn't. Nothing in you is working to drive joy to zero.

The Slime Mold answer is that happiness isn't an emotion at all.

> "Emotions are easy to identify because they are errors in a control system. Like any error in a control system, successful behavior drives the error to zero. This means that happiness is not an emotion." (Slime Mold Time Mold, 2025b)

What happiness is, on their account, is what it feels like when a governor's error gets corrected: "Happiness is what happens when a thirsty person drinks, when a tired person rests, when a frightened person reaches safety." Correct a large error, or correct one quickly, and you get more of it than from a slow incremental fix. That matches ordinary experience better than it has any right to. The best meal of your life was eaten hungry.

Two further consequences follow, and both are testable in principle.

**There is no such thing as unhappiness.** Happiness cannot go negative, because it is not a regulated variable with a set point. What gets called unhappiness is an uncorrected error somewhere in the system, which is a different thing wearing the same word.

**Happiness has a job.** They propose it as a learning and allocation signal: it marks which behaviors worked so the system repeats them, and helps calibrate how much to explore versus exploit. Notably, no governor votes *for* happiness, which is their explanation for why people demonstrably fail to maximize it.

**Where the account runs out.** It handles joy, relief, satisfaction and pleasure by collapsing them into one undifferentiated correction signal. Their words: "there's only one way to feel good," and "all of our words for positive emotion, joy, excitement, pride, are really referring to the same thing, just in different contexts." That is asserted rather than argued, and it leaves amusement, awe and aesthetic pleasure entirely unaddressed. Across the full fourteen-post series, none of those three words appears.

They are straighter about curiosity, which they flag themselves as unresolved: "acting on your fear should make you less afraid, acting on your thirst should make you less thirsty, but acting on your curiosity often seems to make you more curious." A signal that grows

when acted on runs backwards from every other loop in the model, and they list it, along with happiness and surprise, as a non-error signal requiring future work.

**The position taken in this document is narrower than theirs.** Error signals account for the emotions that push an organism toward a target. Happiness is a reasonable reading of what arriving feels like. Amusement, awe and curiosity are unaccounted for, and leaving them unaccounted for is preferable to stretching the framework to cover them, which is precisely the flexibility objection raised in Section 8.

**Status: their claim, clearly stated and repeated across five posts. Not tested.** No one has measured whether reported positive affect scales with the magnitude or rate of error correction, which is what the account predicts.

## 2.4 The real intellectual lineage: credit where due

Almost none of the structural content above is new, and a reader coming to it fresh in 2025 deserves to know that.

James Clerk Maxwell wrote the first mathematical treatment in 1868, in a three-page paper to the Royal Society titled, with beautiful economy, "On Governors" (Maxwell, 1868). Steam engines fitted with centrifugal governors were prone to hunting, surging and sagging instead of holding speed, and Maxwell showed that whether a governor settled or oscillated depended on the relationship between how hard it corrected and how quickly it responded. Gain and damping, in 1868, in the context of machinery. Every oscillating clinical system you'll meet in this document is a descendant of Maxwell's hunting engine.

Norbert Wiener named the field.[1] *Cybernetics: or Control and Communication in the Animal and the Machine* (1948) took the mathematics of feedback control, which had matured enormously during the war in the design of anti-aircraft fire-control systems, and proposed that the same mathematics described nervous systems. That is the direct ancestor of the guidance and control work I trained in, and it began as a problem about aiming at a moving airplane. Five years earlier, Rosenblueth, Wiener, and Bigelow had already argued that purpose in behavior could be defined without mysticism as negative feedback toward a goal state (Rosenblueth, Wiener, & Bigelow, 1943). Wiener's most striking clinical claim was about the cerebellum. He noticed that intention tremor, the overshoot-and-oscillate that appears when a patient with cerebellar damage reaches for a glass, looks exactly like an engineering control loop with its damping removed. The correction is applied too hard and too late, the hand sails past the target, the error reverses, and the limb hunts. Call it what an engineer would call it: a control failure, in the literal sense of the phrase.

Then William T. Powers. In 1973 Powers published *Behavior: The Control of Perception*, and in the same year a paper in *Science* titled "Feedback: Beyond Behaviorism" (Powers, 1973a,

---

[1] Wiener's vocabulary escaped into the popular market fast. Maxwell Maltz's *Psycho-Cybernetics* (1960) borrowed it for a self-help program and sold more than thirty million copies, which is why the word still sounds like a paperback to many clinicians. Appendix A works through where that book was right, where it was missing an estimator, and how it became the source of the false twenty-one-day habit rule.

1973b), which was followed by exactly the kind of argument in the correspondence pages that the title invites. Powers's thesis, developed further in *Psychological Review* five years later (Powers, 1978), was that organisms don't control their behavior. They control their perceptions, and behavior is the variable means by which they do it. He built the model hierarchically, so that higher-level control systems act by setting the reference values of lower-level ones, and a goal at one level is a set point at the level below. He insisted that emotion and physiological drive belonged inside the loop rather than beside it. And he pointed out the two-comparator constraint on bidirectional control that reappears, quoted directly, in *The Mind in the Wheel*.

The content of the 2025 serial is substantially a rediscovery of Powers. That's no criticism. Rediscoveries happen when an idea was right and got buried, and the writing in *The Mind in the Wheel* is clearer and more clinically vivid than most of the Perceptual Control Theory literature, which is part of why the older work stayed buried. But it changes who gets the credit, and it changes what a clinician should read next. There's a fifty-year literature here. Carver and Scheier brought control-theoretic thinking into personality, clinical, and health psychology in 1982 (Carver & Scheier, 1982). Marken and Mansell have kept the Powers program running and have argued its case as a unifying framework (Marken & Mansell, 2013; Mansell & Marken, 2015). Method of Levels is a therapy built explicitly on it (Higginson, Mansell, & Wood, 2011). None of that appears in the serial.

Maxwell in 1868, Wiener in 1948, Powers in 1973, a blog in 2025. The same character keeps walking back on stage the way Ed Baldwin keeps walking back on stage across four decades of *For All Mankind*, older each time, running the same play, and the room acting like it has never seen him before.

Two further absences are worth flagging. W. Ross Ashby, whose *Design for a Brain* (1952) worked out how a control system could find its own set points through ultrastability, isn't cited. Neither is Kenneth Craik, whose *The Nature of Explanation* (1943) proposed that nervous systems build small-scale models of reality and run them ahead of events, the idea that becomes Section 3 of this document. Across every installment of the serial, neither name appears. The gap matters, because Ashby and Craik are precisely the two authors who address the questions the control framework leaves open.

## 2.5 Problems with the control model

The framework is genuinely useful and it's considerably weaker than its presentation suggests. Seven problems, in ascending order of how much they should worry you.

**It's self-published and has essentially no new data.** *The Mind in the Wheel* is a blog serial. It hasn't been through peer review, and its authors are anonymous, which also means nobody can assess whether the reading of the primary literature is competent. More importantly, it reports no new experiments. Its strongest single piece of empirical support is a 1968 study in which Paul Rozin housed Sprague-Dawley rats with water, a salt-vitamin mix, and a liquid cafeteria of sucrose solution, thirty-percent protein solution, and corn oil. The rats held their protein intake near-constant, and when the protein solution was diluted by half or by three-quarters they drank more of it to compensate, imperfectly at the higher

dilution (Rozin, 1968). That's a real result and it's consistent with a protein governor defending a target. It's also a fifty-eight-year-old rat study being asked to carry a proposed paradigm for human psychology.

**The parliament is underspecified.** Governors, votes, weights, gates, pricing, selectors: that's a large number of free parameters and very few constraints on how they're set. Almost any behavioral finding can be accommodated after the fact by positing an additional governor, adjusting a weight, or invoking a gate threshold. A framework that can absorb any result isn't thereby confirmed by any result. Nothing in the serial specifies how many governors there are, how you'd establish that two apparent drives are one governor rather than two, or what pattern of data would falsify a proposed governor's existence. The authors are candid that this is unfinished work and explicitly ask readers to do it. That candor is to their credit. It doesn't fill the gap.

**Identifying the controlled variable is genuinely hard.** That objection is practical, and it's the place where control theory has always struggled. If behavior is variable in the service of a constant perception, you can't read the controlled variable off the behavior, because the behavior is the part that changes. Powers's answer was the Test for the Controlled Variable: disturb a candidate variable and see whether the organism opposes the disturbance. If it does, you've found something being controlled. If it doesn't, you haven't. Marken has developed model-based versions of the procedure (Marken, 2013, 2014). Note what happens in *The Mind in the Wheel*. The authors quote Powers's statement of the Test at length in Part X, endorse it, and then never run it. Not once, on any of the many drives they propose. The one instrument that would turn the framework into science is displayed in its case and left there.

**Where do set points come from?** A control system defends a reference value, and the framework has almost nothing to say about where that value originates, how it's learned, why it differs between individuals, or how it changes. That's no peripheral gap. A great deal of clinical work consists of trying to change what a person is defending: the standard they hold themselves to, the level of closeness they'll tolerate, the degree of certainty they require before acting. If set points are fixed and biological, most therapy is impossible. If they're learnable, the framework owes us an account of the learning, and it doesn't have one. Ashby's ultrastability was an attempt at exactly this problem, which is another reason his absence from the bibliography is unfortunate.

**There's no natural account of belief, meaning, or narrative.** A thermostat has no beliefs. It has a reading and a reference. Extend that to a person and you get a compelling account of drive, motivation, and the arbitration between competing needs, alongside an almost empty account of the material clinicians actually spend their hours on. A patient's conviction that she is fundamentally unlovable is not a set point, not obviously an error signal, and not a sensor reading. It's a proposition with content, connected to other propositions, revisable by evidence and argument, and organized into a story about a life. The control framework can gesture at this by stacking hierarchies of perceptual control until the higher levels are abstract enough to look belief-like, which is what Powers did. Whether that gesture is an explanation or a relabeling is exactly the open question, and the framework as it stands doesn't settle it.

**Cybernetics lost, and the reasons were mixed.** Between the late 1940s and the mid-1970s, cybernetics went from the most exciting idea in the human sciences to a historical curiosity, and cognitive psychology, with its representations, its information processing, and its computer metaphor, took the field. Honesty requires saying that this happened partly for bad reasons. The timing was poor: Powers published in 1973, deep into the cognitive revolution, into an audience with no room for him. The framing was poor: cybernetics arrived wrapped in grand unification claims across biology, engineering, sociology, and management that made careful scientists wince, and the language of "dynamics" and "purpose" carried a whiff of psychoanalysis at exactly the moment academic psychology was working hardest to smell of anything else. Interdisciplinary programs with no departmental home tend to lose funding battles to programs with one.

But it also lost for a good reason, and that shouldn't be dressed up. Over roughly three decades, it under-delivered. The framework generated far more diagrams than experiments, far more reinterpretation of existing findings than new ones, and a very small number of quantitative predictions that anyone went out and tested. Powers's own tracking experiments produced correlations between model and behavior that are genuinely remarkable and that almost nobody outside the tradition has replicated or extended. This is where an engineer's habits are worth borrowing. You estimate from experience, you build the thing, you instrument it, and you let the data outrank the argument you fell in love with. A framework that keeps explaining and rarely predicts eventually loses, and it should. The current revival deserves the same standard.

**Control systems can imitate prediction, which cuts both ways.** Mansell, Gulrez, and Landman argue in *Current Opinion in Behavioral Sciences* that a pure negative-feedback control system, containing no forward model and predicting nothing, will nonetheless generate behavior that an outside observer confidently describes as anticipatory (Mansell, Gulrez, & Landman, 2025). Their title is "The prediction illusion." An organism that keeps a variable near its reference will appear to have foreseen the disturbance it was in fact merely opposing, because the correction begins before the observer notices the perturbation. If they're right, a substantial body of evidence taken to demonstrate a predictive brain is equally consistent with a controlling one.

Read that carefully in both directions before you use it. It undercuts confident predictive-processing interpretations of behavioral data, and it undercuts confident control-theoretic ones by exactly the same logic, because if the two architectures generate observationally equivalent behavior, no amount of that behavior will tell you which one you're looking at. Take what Mansell and colleagues found as a warning rather than a win for control theory. The discriminating experiments have largely not been done, and a clinician choosing between these two accounts on the basis of the published behavioral evidence is choosing on other grounds.

That's the honest position going into Section 3. The control model gives you a precise vocabulary for regulation, a real taxonomy of failure, and a serious argument that symptom-cluster diagnosis conceals mechanism. It doesn't give you belief, it doesn't give you learning, it doesn't give you set points, and it hasn't yet earned the empirical standing its more confident advocates claim for it.

# 3. The Mind as a Prediction Machine

## 3.1 Perception as inference

Close your left eye. Hold your right thumb out at arm's length and look at it steadily. Now move the thumb slowly to the right, fifteen or twenty degrees, keeping your gaze fixed straight ahead. At some point your thumb disappears. It doesn't blur. It's gone. Move it a little further and it comes back.

That's the blind spot, the patch of retina where the optic nerve leaves the eye and there are no photoreceptors at all. Every eye has one, roughly the size of nine full moons laid side by side, and you have never noticed it.

The absence isn't the interesting part. What fills it is.

You don't perceive a black disc, or a gray smudge, or a hole. If your thumb vanishes against striped wallpaper, you see continuous stripes. The visual system never reports missing data. It supplies the most likely answer and hands that answer over with no marking to separate it from anything else you're seeing.

The blind spot is the demonstration everybody knows. The pattern it reveals is everywhere once you look for it.

> *Aside... if you have read Plato's The Allegory of the Cave, this is kind of congruent with that... but enough intellectualizing here....*

*Figure 3.1. The allegory of the cave. The prisoners are chained facing a wall, a fire burns behind them, and carried objects throw shadows onto the wall in front of them. What they see are the shadows, not the objects casting them, and they take the shadows for the world.*

The congruence is closer than a loose analogy. The prisoners are not making an error of reasoning. Their inference is correct given the evidence reaching them, and the evidence reaching them is a projection rather than the thing itself. That is the predictive account of perception in a sentence, arrived at roughly twenty-four centuries early. What the modern version adds is the machinery: the shadows are the sensory input, the prior is what the prisoner already expects a shadow to mean, and precision is how much authority the shadow gets against that expectation. Plato's prisoner freed and turned toward the fire is a prior being violated hard enough to update, which is why it hurts and why he resists it.

You're at a party, loud enough that you can't follow the conversation two feet away, and somebody across the room says your name, and you hear it perfectly. The signal carrying your name was no louder and no cleaner than the noise that swallowed everything else. What made it audible is that your name is a hypothesis your auditory system holds ready at all times, so a fragment of degraded evidence is enough to trigger it. The corollary comes free: you sometimes hear your name when nobody said it. Same mechanism, running on too little evidence.

You have walked down a staircase in the dark, reached the bottom, and taken one more step that wasn't there. The jolt is disproportionate and unmistakable, and nothing hit you. Your motor system issued a prediction (floor, at this height, in one hundred milliseconds) and the prediction was violated. The lurch you felt was the size of the mismatch.

And then there's the dress. In 2015 a photograph of a striped garment divided the internet into people who saw it as white and gold and people who saw it as blue and black, each group finding the other's report literally unbelievable. Lafer-Sousa, Hermann, and Conway measured the split and located its source: the image is genuinely ambiguous about the color of the light falling on the dress, and observers resolve that ambiguity using an implicit assumption about the illumination (Lafer-Sousa, Hermann, & Conway, 2015). Assume warm indoor light and the brain discounts yellow, leaving blue and black; assume cool daylight and it discounts blue, leaving white and gold. Same photons, same retinas, different assumption, different color. Not different opinions about the color. Different color, seen.

Hermann von Helmholtz named this in the 1860s and nobody has improved on his term: **unconscious inference**. The eye receives a two-dimensional pattern of light intensities, and such a pattern is compatible with an infinite number of three-dimensional scenes. Something has to choose. The visual system chooses by combining the sensory evidence with what it already knows about how the world tends to be arranged (light comes from above, surfaces are usually opaque, objects tend to be rigid) and the choosing happens beneath awareness, fast, with no experience of having chosen. What reaches you is the conclusion, delivered as though it were the premise.

The claim to carry forward is this: **perception is a construction. It's the brain's best guess about the causes of its sensory input, constrained by the evidence but not determined by it.** Everything you see, hear, and feel is a hypothesis the data haven't yet overturned.

I want to flag something here, because it runs under the rest of this section and it comes back hard in 3.6. I spent years as an engineer before I ever sat with a client, and the finding that took me longest to accept is that people don't reason their way to a position and then develop feelings about it. They decide emotionally and then manufacture a logical-sounding case for what they already chose. A Vulcan couldn't run a real business, because rationalization runs in front of logic in every room I've worked in, including the ones full of very smart people. What Helmholtz described is that same order of operations one level down, in the wiring. The guess goes first. The evidence gets a vote, not a veto.

One discipline corrects for it, and I've watched it work on an airframe and in a treatment plan. Estimate from experience, TEST, evaluate the result without flinching, and let the data outrank the argument you fell in love with. Hold that standard. In 3.6 we apply it to this framework itself.

## 3.2 Prediction error

If the brain is generating guesses, something has to keep the guesses honest. That something is the mismatch.

The standard formulation reverses the flow you were probably taught. In the textbook picture, information enters at the sensory surface and travels upward through successively more abstract stages until it becomes perception at the top. In the predictive picture, the dominant traffic runs the other way. Higher regions send **predictions downward**, a running account of what the level below should be receiving, given the current hypothesis about what's out there. Lower regions compare the prediction against what they actually got, and send **only the difference upward**. What ascends is the part of the signal nobody anticipated: the **prediction error**.

Rao and Ballard gave this its influential neural formulation in 1999. They built a hierarchical model of visual cortex in which feedback connections carried predictions and feedforward connections carried residual error, trained it on natural images, and found it reproduced a set of otherwise puzzling findings, in particular end-stopping and other extra-classical receptive field effects, where a neuron's response to a stimulus in its receptive field is suppressed by context outside it (Rao & Ballard, 1999). Their explanation was direct: the surrounding context lets the higher level predict what the neuron should be seeing, the prediction is subtracted, and the neuron falls silent because it has nothing new to report. Context suppression, on this reading, is what successful prediction looks like from the outside.

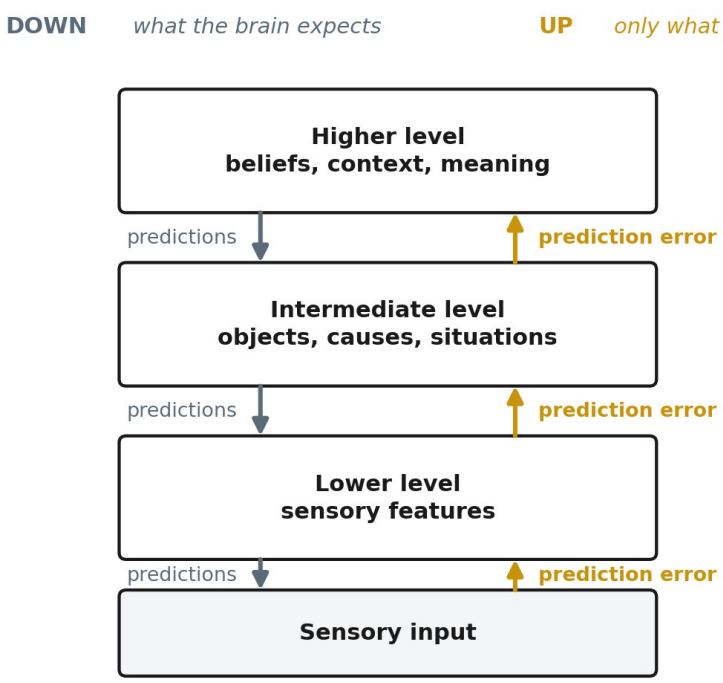


*Figure 3.2. Predictions descend, prediction errors ascend.*

The efficiency argument is easy to feel. Most of what strikes your senses in any given second is exactly what struck them the second before, and transmitting only the departures from expectation compresses the traffic enormously, with the useful side effect that whatever does get passed along is automatically the part worth attending to.

The word **surprise** needs pinning down here, because the technical sense is narrower than the everyday one. Surprise in this literature is a quantity: how improbable the sensory input was under the model the system currently holds. It has nothing to do with the feeling of astonishment. Something can be highly surprising in that sense while producing no conscious experience of surprise at all, and much of what your brain treats as surprising never reaches awareness. Run it the other way and you can feel startled by something your visual system predicted perfectly well. Keep the two apart.

Clinicians already know a version of this idea, though probably not under this name.

In 1997, Schultz, Dayan, and Montague published one of the most cited papers in behavioral neuroscience (Schultz, Dayan, & Montague, 1997). Recording from midbrain dopamine neurons in monkeys learning to associate a cue with a juice reward, they found a pattern that had confused the field. Early in training, the neurons fired to the juice. After learning, they stopped firing to the juice and fired instead to the cue that predicted it. And if the cue appeared and the juice didn't arrive, the neurons went silent at exactly the moment the juice was due, a dip below baseline, precisely timed.

What dopamine tracks is the difference between the reward received and the reward expected: **reward prediction error**. An unexpected reward produces a burst; a fully predicted reward produces nothing, because nothing was learned; an expected reward that

fails to arrive produces a negative signal. The neurons were computing a subtraction, and the quantity they computed was formally the same as the error term Rescorla and Wagner had inferred from animal learning data a quarter-century earlier (Rescorla, 1972).

This matters here for one reason. Most clinicians already accept the reward prediction error account, because it underpins how we talk about reinforcement, addiction, anhedonia, and behavioral activation. That account is the same idea as predictive coding, applied to value instead of sensation. If you already believe the brain learns about reward by tracking the gap between expectation and outcome, you've already accepted the architecture. The predictive processing claim is simply that this is the operating principle of the cortex and not a special-purpose reward circuit.

### 3.3 Precision: the concept the whole document turns on

Everything so far can be stated without much subtlety. This next idea can't, and it's the one that does the clinical work.

You're trying to identify a bird, and someone hands you a photograph. If the photograph is sharp, every feather, the exact shade of the throat patch, the shape of the bill, then whatever you thought the bird was before hardly matters. The evidence is good and it should dominate. If the photograph is a smeared gray blur taken through a rainy window, it should barely move you at all; you should mostly go on where you were standing, what season it is, and what birds are common there. The blurry photograph still counts as evidence. It counts as *weak* evidence, and the rational move is to weight it accordingly.

That weighting is what the field calls **precision**. Precision is the system's estimate of how reliable a signal is, meaning how much to trust it. Formally it's the inverse of expected variance. Practically it's confidence, and it's assigned separately to predictions and to incoming evidence.

Four terms carry the entire machinery. In plain language:

**Prior:** what the system expected before any evidence arrived.

**Likelihood:** how well the incoming evidence fits a given hypothesis.

**Prediction error:** the part of the input the prediction failed to cover.

**Precision:** how much the system trusts a signal, assigned separately to the prediction and to the evidence.

Here's the relation the rest of this document depends on, stated in words rather than symbols:

> **What you end up believing is a weighted blend of what you expected and what you sensed, and the weights are how much you trust each one.** Trust the prediction more than the evidence, and perception drifts toward expectation. Trust the evidence more than the prediction, and perception is dragged toward the input.

Two things follow immediately, and both should be uncomfortable.

The first is that identical evidence produces different perceptions in different people, without either of them being irrational, if they assign different precision to it. The second is that a system can generate a false percept in two entirely different ways: by holding its prior too confidently, or by treating its sensory evidence as too unreliable. Behaviorally these can look the same. We come back to that in 3.6, because it's a deeper problem than it first appears.


*Figure 3.3. Precision weighting: the same evidence, three different levels of trust.*

The move that made this concept indispensable was Feldman and Friston's proposal that **attention just is the assignment of precision to sensory channels** (Feldman & Friston, 2010). When you attend to something, on this account, you aren't shining a spotlight on it or moving a limited resource around. You're raising your estimate of how reliable that channel is, which raises the gain on its prediction errors, which lets it exert more influence over what you end up perceiving. Feldman and Friston showed that a model doing this reproduced classic attention findings, the Posner cueing effect with its costs and benefits, and the pattern in the Posner paradigm where invalidly cued targets are responded to slowly, without any separate attention module.

Attention becomes a control parameter rather than a faculty. Something that felt like a psychological capacity turns into a setting on a signal-weighting mechanism, and settings can be miscalibrated in ways faculties cannot.

The obvious next question is what implements precision in the brain, and the honest answer is that the mapping is loose. The usual candidates are the neuromodulators. Dopamine is the most frequently named, generally cast as encoding the precision of beliefs about policies or the reliability of higher-level expectations. Acetylcholine is often assigned the precision of sensory evidence, and noradrenaline the estimate of environmental volatility, meaning how quickly the world is changing and therefore how much old information should be discarded. That division traces back to Yu and Dayan's distinction between expected and unexpected uncertainty (Yu & Dayan, 2005) and has been extended within the active inference literature (Parr & Friston, 2017).

Treat this scheme with real caution. Each of these transmitters has many functions across many systems, the assignments differ between authors, and the evidence is largely indirect:

pharmacological manipulations with broad effects, computational fits to behavior, and inference from lesion and disease states. Nothing in the literature approaches a demonstration that a specific neuromodulator implements a specific precision parameter in a specific circuit. The mapping is a working hypothesis presented, in a good deal of that literature, with more confidence than it has earned. When you read that a drug "increases sensory precision," read it as a model-fitting result and NOT as a measurement.

## 3.4 Active inference and the body

There are two ways to reduce a mismatch between what you predicted and what you sensed. You can change the prediction. Or you can change the world so the prediction comes true.

The second option is action, and treating it that way is the step from predictive coding to **active inference**. On this account, no separate motor system receives commands and executes them. Movement is the fulfillment of a proprioceptive prediction. The brain predicts the pattern of muscle-spindle and joint signals that would occur if the arm were already extended, that prediction generates proprioceptive error, and the classical reflex arcs at the spinal level resolve the error the only way available to them, by moving the arm until the predicted sensations arrive. The prediction functions as a specification of what should be sensed, and the body makes it true.

Notice that this is structurally the same claim Powers made about controlling perception, arrived at from a different direction. Section 2 and Section 3 are closer relatives than their vocabularies suggest, and Section 4 takes that seriously.

One requirement falls out immediately. For a proprioceptive prediction to drive movement rather than be corrected by the evidence, the incoming proprioceptive error has to be temporarily down-weighted, its precision reduced, or the system would simply conclude the arm hasn't moved and revise the prediction instead. This is **sensory attenuation**, and it's why you can't tickle yourself: self-generated sensation is predicted and therefore attenuated (Brown, Adams, Pareés, Edwards, & Friston, 2013). Hold on to it. It's the hinge of the best clinical application in 3.5.

Turn the same machinery inward and it becomes clinically unavoidable. The brain receives a continuous stream from inside the body: cardiac, respiratory, gastric, vascular, immune, metabolic. If perception in general is inference, then perception of the body's internal state is inference too. Seth's proposal is that emotional experience is **interoceptive inference**, your best hypothesis about the causes of your internal signals, generated by the same machinery that produces your best hypothesis about the causes of light on the retina (Seth, 2013; Seth & Friston, 2016). A racing heart is data. What it means (fear, excitement, exertion, caffeine, illness) is a conclusion, arrived at inferentially, using context and prior expectation, and delivered to you as a feeling instead of as a judgment. It's WebMD implemented in hardware: one real signal in, the most confident available diagnosis out, and no confidence interval printed anywhere on it.

Barrett and Simmons made the anatomical case, arguing that the agranular visceromotor cortices, anterior insula and anterior cingulate, issue interoceptive predictions against

which the ascending signals are compared, rather than receiving interoceptive information and reacting to it (Barrett & Simmons, 2015). The direction of traffic is the substantive claim: the brain anticipates the body and notes the residuals rather than primarily reading it. Barrett's theory of constructed emotion builds on this (Barrett, 2017), as does the active-inference account of allostasis and interoception in depression (Barrett, Quigley, & Hamilton, 2016).

Peter Sterling's argument, stated most compactly in a 2012 paper, is that homeostasis is the wrong model for how a body is actually governed (Sterling, 2012). Homeostasis is reactive: a variable deviates, the deviation is detected, a correction follows. **Allostasis** is predictive: the system anticipates demand and adjusts *before* the deviation occurs. Blood pressure rises in the seconds before you stand. Insulin is released at the sight and smell of food, not in response to the glucose. Cortisol climbs in the hour before waking. A purely reactive regulator would be permanently behind, so efficient regulation means predicting demand and being wrong as rarely as possible. Recall the thirst circuit from Section 2, which switches off during swallowing rather than after absorption, allostasis running inside a system that looks, on a diagram, purely homeostatic.

The clinical implication isn't decorative. If regulation is predictive, chronic dysregulation needs no broken sensor and no failed correction. It can consist entirely of a body being prepared, accurately and efficiently, for demands that aren't coming. A system still expecting a world it no longer lives in.

## 3.5 The clinical versions

The framework has been applied to almost every psychiatric presentation. The applications vary enormously in how well they're supported, and it's worth being explicit about which is which.

**Psychosis as overweighted priors.** The strongest single experiment is Powers, Mathys, and Corlett's *Science* paper (Powers, Mathys, & Corlett, 2017). They took four groups (voice-hearers with a psychosis diagnosis, voice-hearers without one who identified as psychics, people with psychosis who don't hallucinate, and healthy controls) and conditioned them all to associate a visual cue with a faint tone, then presented the cue with no tone. Both voice-hearing groups reported hearing the tone, with high confidence, far more often than the other two. Computational modeling located the difference in the weighting of prior expectations relative to sensory evidence, and the effect scaled with clinical hallucination severity. That's close to a demonstration that a conditioned prior can manufacture a percept in the absence of the stimulus. Sterzer and colleagues review the wider predictive-coding account of psychosis, including the difficulty that different stages of illness appear to require opposite claims about whether priors are too strong or too weak (Sterzer et al., 2018; see also Corlett et al., 2019; Adams, Stephan, Brown, Frith, & Friston, 2013).

**Functional neurological disorder.** This is the framework's best clinical case, because it solves a problem nothing else solved. Edwards, Adams, Brown, Pareés, and Friston proposed that in FND a strongly held prior about the body, a belief in a paralyzed limb or

an expectation of a tremor, is granted such high precision that it dominates the actual sensory evidence, while attention to the affected limb further alters the precision of incoming signals (Edwards, Adams, Brown, Pareés, & Friston, 2012). The symptom is genuinely involuntary from the patient's side, because the prediction operates at a level the person has no access to, *and* it's generated by the patient's own brain rather than by a lesion. That's exactly the combination clinicians observe and that a century of theorizing failed to accommodate without implying malingering or repression. It also makes testable predictions: sensory attenuation should be abnormal in these patients, and it is (Pareés et al., 2014). FND is where predictive processing has done the most clinical good.

**Chronic pain and placebo.** Büchel, Geuter, Sprenger, and Eippert set out placebo analgesia as a predictive coding phenomenon: the experienced intensity of pain is a weighted combination of nociceptive input and expectation, so an expectation of relief, held with sufficient precision, changes not the report of the pain but the pain (Büchel, Geuter, Sprenger, & Eippert, 2014). That dissolves a persistent confusion about whether placebo effects are "real." On this account no separate real pain sits behind the experienced one. There's an inference, and expectation is one input to it. The extension to chronic pain, that pain can persist as a highly precise prior long after the nociceptive evidence has faded, is coherent and clinically appealing, and considerably less directly evidenced than the placebo work.

**Anxiety and volatility.** Browning, Behrens, Jocham, O'Reilly, and Bishop had participants learn which of two options predicted an electric shock, in an environment whose statistics shifted periodically (Browning, Behrens, Jocham, O'Reilly, & Bishop, 2015). Healthy low-anxious participants adjusted their learning rate appropriately: fast updating when the world was changing, slow updating when it was stable. High trait-anxious participants failed to make that adjustment, learning at a similar rate regardless. The finding reframes anxiety as a failure to estimate how much the world is changing, and therefore a failure to know how much any given piece of bad news should count. The study is well designed and has been influential. It's also one experiment with a modest sample, and the trait-anxiety effect is a correlation in a non-clinical population.

**Fatigue and depression.** Stephan and colleagues proposed **allostatic self-efficacy**: fatigue as the signal of persistently unresolved interoceptive prediction error, and depression as what follows when the system arrives at a low-level belief that its own regulatory actions no longer work (Stephan et al., 2016). Depression, on this reading, is an inference about regulatory competence sitting underneath the mood. It connects fatigue, inflammation, interoception, and mood in one structure, and it should be read as a theoretical proposal awaiting direct test rather than an established finding. The same caution applies to Petzschner and colleagues' broader computational-psychosomatics program (Petzschner, Weber, Gard, & Stephan, 2017).

## 3.6 Problems with the predictive model

The predictive framework is the most influential idea in theoretical neuroscience of the last twenty years. It also has serious problems, and clinicians are rarely told about them.

Go back to the order of operations from 3.1. Decide first, then build the case. That happens to fields, not only to people, and this is a field that fell in love with an elegant argument and then mostly stopped trying to break it.

Psychosis. Chronic pain. Placebo. Anxiety. Functional neurological disorder. Fatigue. Depression. Autism. Addiction…

Applied to all of that. Tested hard against a small fraction of it. Weigh the following before leaning on any of it.

**The unfalsifiability charge.** Bowers and Davis put the objection most sharply in *Psychological Bulletin* under the title "Bayesian just-so stories in psychology and neuroscience" (Bowers & Davis, 2012). A Bayesian model has three components a modeler can adjust, the prior, the likelihood function, and the utility or cost function, and each is typically inferred from the data rather than measured independently. Given that freedom, a model can be fitted to almost any pattern of behavior, including patterns that contradict each other. Their taxonomy distinguishes three kinds of objection. The **extreme** claim is that Bayesian theories are unfalsifiable in principle. The **methodological** claim is that in practice the free parameters get chosen after seeing the data, so the fit demonstrates flexibility rather than truth. The **theoretical** claim is that showing behavior is *consistent with* optimal inference tells you little, because many non-Bayesian mechanisms produce near-optimal behavior, and the demonstration therefore doesn't identify a mechanism. Bowers and Davis endorse the second and third rather than the first, and that's the version clinicians should hold. The framework isn't unfalsifiable. A great deal of the published work simply hasn't tried very hard to falsify it.

**Unification by fiat.** Litwin and Miłkowski's critique in *Cognitive Science* argues that predictive processing has expanded by absorption rather than by explanation (Litwin & Miłkowski, 2020). Perception was restated in predictive terms. Then action. Then attention. Then learning. Then emotion, interoception, and psychopathology. Each restatement was announced as unification, but a redescription only unifies if it generates constraints, and here it frequently generates none. Their charge is that the framework's development has been arrested: the vocabulary keeps growing while the mechanistic detail does not, and the theory is protected from disconfirmation by the ease with which any result can be restated in its terms. A theory that grows after every disconfirmation is doing something other than science.

**The identifiability problem.** This is the most concrete of the objections and the one with the sharpest clinical consequence. Recall from 3.3 that the outcome is a weighted blend of prior and evidence. Now consider two people: one with a weak, uncertain prior and highly reliable sensory evidence, the other with a strong, confident prior and unreliable evidence. Under the right numbers, these two produce **identical** perceptual outcomes. Nothing in the behavior distinguishes them. The arithmetic guarantees it, which makes this structural and not a technical inconvenience you can engineer around. It means a claim of the form "in this disorder, priors are too strong" is usually not separable, on the available behavioral data, from "in this disorder, sensory precision is too low." Those two have completely different treatment implications. The psychosis literature shows the problem in the open: strong-

prior accounts and weak-prior accounts have both been advanced, sometimes for the same phenomena at different stages of illness, and the datasets frequently can't adjudicate. Independent measurement of one term (through pharmacology, through neurophysiology, through experimental manipulation of actual stimulus reliability) is the only way out, and it's done far less often than it should be.

**The dark room problem.** If a brain is built to minimize prediction error, why doesn't it find a dark, silent, unchanging room and stay there forever? Sensory input in that room is maximally predictable and error would fall to nearly nothing. Wednesday Addams would call that a good weekend.

Friston, Thornton, and Clark gave the standard reply: an organism's model includes expectations about itself, that it will eat, move, socialize, encounter novelty, and a dark room violates those expectations catastrophically, generating enormous error of exactly the kind that gets the organism out of the room (Friston, Thornton, & Clark, 2012). Sun and Firestone pressed on this in *Trends in Cognitive Sciences*, arguing that the reply either smuggles in the very drives it was supposed to explain, or renders the principle vacuous by allowing any behavior at all to be described as error minimization given a suitably chosen model (Sun & Firestone, 2020). The exchange produced two separate replies in the same issue, Van de Cruys, Friston, and Clark on one line of defense, Seth, Millidge, Buckley, and Tschantz on another (Van de Cruys, Friston, & Clark, 2020; Seth, Millidge, Buckley, & Tschantz, 2020). Anybody who has made the second trip to Home Depot for the same repair knows what it means when the first answer needs a follow-up answer. Two defenses, not entirely compatible with each other, against a two-page critique. The problem isn't fatal and it hasn't been cleanly resolved.

**Four different claims wearing one name.** Much of the confusion in the clinical literature comes from a conflation that's easy to fix once you see it. At LEAST four separate propositions travel together under the predictive banner, and they differ enormously in how testable they are.

*Predictive coding as a neural algorithm.* The claim that cortical circuits literally compute and transmit prediction errors, with specific cell types and laminar connections doing specific jobs. This is a concrete neurophysiological hypothesis, it's testable, and it's being tested.

*The Bayesian brain hypothesis.* The claim that the nervous system represents and combines information in an approximately Bayes-optimal way. Testable at the behavioral level, subject to the identifiability problem above, and it specifies no mechanism by itself (Colombo & Seriès, 2012).

*Predictive processing as a cognitive architecture.* The claim that perception, action, attention, learning, and emotion are all instances of hierarchical prediction error minimization (Clark, 2013; Hohwy, 2016). This is a framework-level proposal. It gets evaluated on coherence and productivity more than on any single experiment, and it's where the Litwin and Miłkowski critique bites hardest.

*The free energy principle.* The claim that any system maintaining its boundaries against dissipation must, of mathematical necessity, appear to minimize variational free energy (Friston, 2010). This is a formal result about the class of systems that persist. Whether it's empirically falsifiable at all is genuinely disputed, including by people sympathetic to it.

A study supporting the first says almost nothing about the fourth. Papers routinely cite across these levels as though they were one body of evidence. They are not.

**Pathological precision is stipulated, not derived.** Crippen's critical review in *Behavioral Sciences* makes an argument that ought to trouble anyone using this framework clinically (Crippen, 2026). Predictive accounts of psychopathology typically proceed by identifying a symptom, positing a precision abnormality that would produce it, and then presenting the symptom as evidence for the abnormality. But the framework contains no independent standard for what the *correct* precision would have been. Crippen's title, "Errors or Adaptations?", is the point: a precision setting that produces distress in one environment may be well-calibrated to another, including the environment in which the person actually learned it. Without an independent criterion, the designation "pathological" is imported from clinical judgment, dressed in computational terms, and then read back out as though the model had derived it. That doesn't make the accounts wrong. It means they're frequently less explanatory than they appear, and clinicians should notice when ~~a model has just restated the diagnosis in equations~~ a model's conclusion was already contained in how it was set up.

**The direct neural evidence is narrower than advertised.** Keller and Mrsic-Flogel's influential review argues that predictive processing is a canonical cortical computation and assembles the supporting physiology, including mouse visual cortex neurons that respond to mismatches between predicted and actual visual flow during locomotion (Keller & Mrsic-Flogel, 2018). That work is real and it's good. It's also a specific set of findings in a small number of systems, mouse V1 and some auditory paradigms, that gets cited regularly to support claims about a general cortical principle.

Furutachi and Hofer's recent review in *Annual Review of Neuroscience* is the skeptical companion piece and should be read alongside it (Furutachi & Hofer, 2026). Their central objection is **signal underdetermination**: a neuron whose firing correlates with the difference between expectation and input hasn't thereby been shown to be computing a prediction error. The same response profile can be produced by adaptation, by stimulus-specific fatigue, by attention-related gain changes, by novelty detection, or by ordinary nonlinear tuning to conjunctions of stimulus and behavioral state. Distinguishing a genuine prediction-error neuron from those alternatives requires experimental designs most published studies haven't used. They don't conclude that cortical prediction errors are absent. Their conclusion is that the confidence with which the field asserts them exceeds what the recordings currently establish.

That's the state of the evidence going into Section 4. The predictive framework offers something the control framework does not: an account of belief, of learning, of how a system's model of the world shapes what it perceives, and a single parameter, precision, that connects attention, confidence, symptom formation, and therapeutic change. It also

carries a persistent identifiability problem it hasn't solved, a habit of describing rather than predicting, a physiological evidence base narrower than its rhetoric, and a tendency to stipulate the pathology it claims to explain.

Estimate, test, evaluate honestly, revise. That's the discipline this framework describes when it describes a brain. It's also the discipline the framework itself has largely been spared. Both of these frameworks are more useful than either is proven.

---

# 4. Combining the models: a learning theory of emotional regulation

## 4.1 The join

The two frameworks in the last two sections describe the same organism and barely acknowledge each other. Read them carefully and they stop looking like competing accounts of one thing. They're accounts of two adjacent things, and you can see the seam the moment you ask a simple question about the thermostat.

A thermostat on your wall has a sensor that reads a number, and it takes that number at face value. Nothing in the device estimates anything. All of its intelligence sits in the rule connecting the reading to the set point.

No biological system gets that deal.

Its sensors are noisy, slow, ambiguous, and, decisively, they measure proxies instead of the thing being regulated. Plasma osmolality is not thirst. Firing rates in thermoreceptors are not core temperature. A racing heart is not danger. In every case the system has to infer the state of the variable it cares about from evidence that underdetermines it, continuously, in real time, while acting.

That inference problem is exactly what predictive processing is a theory of.

There's the join. Predictive processing describes how the estimate gets made. Control theory describes what happens to the error once the estimate exists. The cybernetic account has a sensor-shaped hole in it, and predictive processing is the right shape to fill it.

None of that is exotic. It's the architecture of every serious control system built since 1960, where a state estimator and a controller get combined into one working loop (Kalman, 1960; Joseph & Tou, 1961; Wonham, 1968). I spent years around that architecture before I ever sat in a therapy room. The unusual thing isn't the engineering. It's that psychology built both halves in separate literatures that don't cite each other.

Two features of the wiring do real work, and both are easy to draw wrong.

**The estimate becomes the target.** The estimator combines prior with evidence and produces an estimate of the current state. That estimate is simultaneously the updated prediction of what should be happening, and that prediction is what the controller defends. There is no separate goal parameter sitting off to one side. The thing the system predicts IS the set point it acts to maintain.

**The sensory return feeds both halves.** Action changes the world, the world returns sensation, and the same sensation does two jobs at once. It enters the estimator as evidence to be weighed against the prior. It also enters the controller, where it is compared against the set point. One signal, two destinations.

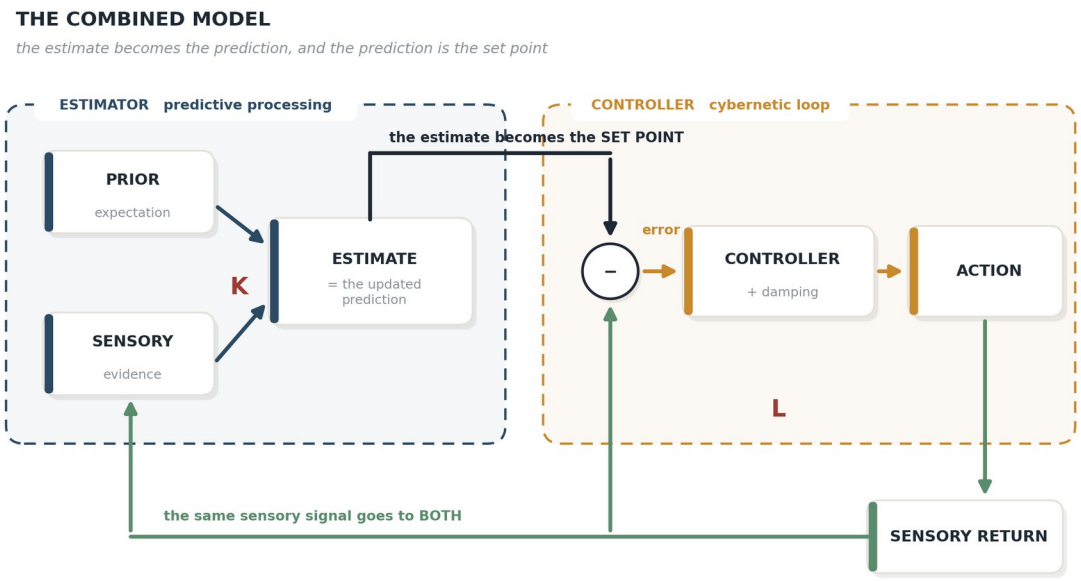


*Figure 4.1. The combined model. The estimate of state becomes the updated prediction, which serves as the set point. The sensory return from action feeds both the controller and the estimator.*

## 4.2 Two gains, and which is which

The combined picture has two quantities in it that both get called "gain." Keeping them apart is the whole practical value of this exercise.

**Observer gain**, written K, sets how far a new piece of sensory evidence moves your estimate of the current state relative to what you already expected. High observer gain means the senses win and the estimate tracks the incoming signal closely, noise included. Low observer gain means expectation wins and the estimate is smooth, stable, and possibly out of date. This is the quantity predictive processing calls **precision**, and it's set by an implicit judgment about relative reliability. How much do I trust this signal, and how much do I trust my model?

**Controller gain**, written L, sets how forcefully the system acts once an error is on the books. It has nothing to do with belief. Two people can land on exactly the same read of a situation and respond with wildly different intensity, and that difference is L.

A third quantity, **damping**, sets how the correction gets delivered over time. Does the system move decisively and settle, or does it overshoot and have to come back? A fourth, **delay**, is the lag between a disturbance and the correction reaching it.

Because the estimate becomes the target, K does something the classical picture hides: it determines **where the target sits**, not merely how accurate the reading is. Move K and you move the goal. Move L and you change how hard the system drives toward a goal that hasn't moved.

The two are in series, in that K's output is the quantity L operates on, and simultaneous, in that both run on every pass of the loop. This isn't a relay. Sensation arrives, the estimate updates, the target moves with it, the error recomputes, the action adjusts, and the cycle repeats without pausing to hand anything off.

That answers a question raised in the margin of an earlier draft: does the two-gain distinction apply to the priors, or to the predictions?

Neither. That's the point.

K IS the quantity that arbitrates between prior and evidence. The entire prior-versus-evidence argument lives inside that one parameter. L sits downstream of the argument entirely. By the time L operates the estimate is settled, and the only question left is what to do about the gap between that estimate and the set point.

The prior-versus-likelihood debate that has occupied the autism literature for fourteen years is a debate about K. It has never had anything to say about L. Which is a reasonable explanation for why it has never explained the repetitive features of the phenotype.

## 4.3 Where emotion sits

If an emotion is the error signal in a behavioral control system, and that's the cybernetic claim, then combining the frameworks changes something worth noticing about it. The error isn't computed against a measured state. It's computed against an **inferred** one.

Anxiety on this reading isn't the registration of danger. It's the gap between an inferred level of danger and the level the system is prepared to tolerate. Loneliness is the gap between inferred social connection and a set point for it. Both quantities on the left of that comparison are estimates, assembled by weighing what you expected against what you sensed.

Sit with that, because it explains something every clinician watches happen and the pure cybernetic account can't handle.

If emotions were error signals computed on measurements, reframing shouldn't touch them. A thermostat doesn't feel differently about 62 degrees once you explain the situation to it. But emotions are famously responsive to reframing, context, and new information, and they're responsive precisely because the thing being compared to the set point is an inference. Inferences move when the evidence moves or the expectation moves.

It also explains the ceiling on reframing, which clinicians watch just as often. Changing a high-level belief doesn't automatically change a low-level estimate, because the estimate is assembled across a hierarchy and the lower levels carry their own precision settings.

"I know it's irrational and I still feel it" isn't a failure of insight. It's what updating one level while another level keeps weighing the evidence exactly as before actually feels like.

## 4.4 A learning theory of regulation

The most useful thing that falls out of combining the frameworks is that the regulatory parameters become **learned quantities**. That's what earns the phrase "a learning theory of emotional regulation," and it carries the clearest clinical implication in this document.

None of K, L, damping, or the set point is fixed at birth in any interesting sense. Each one gets estimated from experience, and each one gets estimated from a statistic of the environment the person actually lived in.

**Observer gain is learned from volatility.** How much you should let new evidence move your estimate depends on how fast the world actually changes. In a stable environment a surprising observation is probably noise and should be discounted. In a volatile one it's probably signal and should move you a long way. Humans demonstrably track this. Learning rates rise under volatility and fall under stability (Behrens, Woolrich, Walton & Rushworth, 2007). Somebody who grew up in a genuinely unpredictable house learned to weight incoming evidence heavily, because in that house it WAS heavily weighted evidence.

**Controller gain is learned from consequence.** How hard you should act on a given error depends on what happened historically when you under-responded. In an environment where small problems reliably became large ones and nobody else stepped in, a high controller gain isn't a malfunction. It's the correct setting for that environment, learned correctly.

**Set points shift by allostasis.** They aren't fixed targets. They're predictively adjusted ones, moved in anticipation of demand rather than in response to it (Sterling, 2012). Stephan and colleagues (2016) formalize this as prior beliefs defining the reference values for homeostatic reflexes, and build a clinical model of fatigue and depression on top of it.

**Damping is learned from overshoot.** A system that has overcorrected repeatedly and paid for it should learn to wait. A system that got punished for waiting learned not to.

Put those four together and you get a claim worth saying out loud, because it reframes a great deal of what walks into a consulting room:

**Most of what presents as dysregulation is a correctly learned parameter running in the wrong environment.**

The settings aren't errors. They're accurate solutions to a previous problem, carried forward into a situation with different statistics. That's a rational-adaptation account of psychopathology instead of a deficit account. It matches what people tell you about their own histories. And it's considerably easier to say to a client than most of the alternatives.

It carries a warning too. A parameter learned over fifteen years from consistent evidence is not going to be revised by one good afternoon of argument. Under this model the rate at which a regulatory parameter can change is itself governed by how much evidence stands behind it. That's why insight is a poor lever and repeated disconfirming experience is a good one.

Which is also where my other profession keeps showing up. You don't re-tune a control system by explaining to it that its gain is wrong. You run it, you instrument it, you look at the data, and you let the results outrank whatever you believed at the whiteboard. People are harder than hardware, and the order of operations is the same.

## 4.5 What oscillation does and doesn't tell you

A second question came in the margin, and it's sharper than it looks. Do the oscillation findings point at a problem with priors, with predictions, or is it inconclusive?

Strictly, on their own, they're inconclusive about priors. That turns out to be informative rather than disappointing.

Oscillation is a controller-side phenomenon. A loop oscillates when the correction is too strong relative to the damping, or when delay makes corrections arrive based on stale error. None of those three quantities, L, damping, delay, is the prior-versus-evidence weighting. The postural and force-tracking findings don't adjudicate the fourteen-year argument about K, and they were never going to.

Which is the argument for taking them seriously. If the repetitive features of the phenotype are a controller-side signature, then a literature that only ever theorized about the observer side would be expected to have failed to explain them.

That's exactly what happened.

Two honest complications. High observer gain also destabilizes a loop, because an estimate chasing sensory noise feeds a jittery reference into the controller, and the controller then chases the jitter. Total loop gain includes both. K and L don't separate cleanly in their effects on stability.

The second complication is more useful, because they leave DIFFERENT signatures. Instability driven by high K should look like broadband noise-tracking: variance scaling with sensor noise, no preferred frequency, improvement when you clean up the sensory signal. Instability driven by high L with low damping should show a resonant peak at a characteristic frequency, largely independent of sensor noise, worsening when you add delay.

Those are distinguishable with a frequency-domain analysis. In several cases, on data somebody has already collected.

That's the discriminating experiment, and it doesn't require recruiting a single new participant.

## 4.6 The phase-locking idea

A third margin question deserves more than a footnote, because it extends the model into the social domain where the autism literature has been weakest. Could we be running a phase-locked loop, and does autism disturb it?

A phase-locked loop regulates the TIMING of an internal oscillator against an external rhythm. Three of its properties are worth borrowing. It has a **capture range**, meaning it can only lock onto inputs close enough to its own natural frequency. It has a **lock range**, wider than the capture range, meaning once locked it can hold on across a broader band than it could have grabbed from a standing start. And when it can't lock at all it doesn't go quiet. It **free-runs at its own natural frequency**.

Human social interaction is full of mutual timing regulation. Conversational turn-taking, gaze exchange, postural convergence, the coordinated rhythms of caregiving. Two systems adjusting to each other's timing, with a substantial empirical literature behind it under the heading of interpersonal or biobehavioral synchrony (Feldman, 2012). Neural entrainment to external rhythm is well documented, and Beker, Foxe and Molholm (2021) found it REDUCED in autistic children.

Three things follow if you take the analogy seriously. Each one is testable.

**Stimming as free-running.** If a regulatory loop can't lock onto external structure, the prediction isn't that it goes silent. It's that it runs at its own natural frequency. That's a mechanistically specific account of self-generated rhythmic behavior, and it predicts something the predictability account doesn't. Stimming should INCREASE when external rhythm is present but unlockable, meaning irregular, unpredictable, or at the wrong tempo, and decrease when external rhythm is either absent entirely or well matched. A pure predictability account predicts things get monotonically better as you add environmental structure. The phase-locking account predicts a non-monotonic curve, where the worst condition is structure that almost works.

**Double empathy, stated formally.** Two oscillators with mismatched natural frequencies can't lock to each other. Two with matched natural frequencies lock easily, whatever those frequencies happen to be. That's a precise restatement of Milton's (2012) double empathy problem, and it makes the same prediction: the difficulty is relational rather than located in one party, and autistic-to-autistic interaction should work better than the mixed case. Crompton and colleagues (2020) found precisely that. Information transfer in autistic-to-autistic chains was as effective as in non-autistic chains, and both beat mixed chains. A capture-range account predicts that result. A social-deficit account doesn't.

**Capture versus lock.** The distinction predicts that starting an interaction should cost more than sustaining one, which matches what clinicians hear constantly. The first five minutes are expensive. Familiar people are cheap.

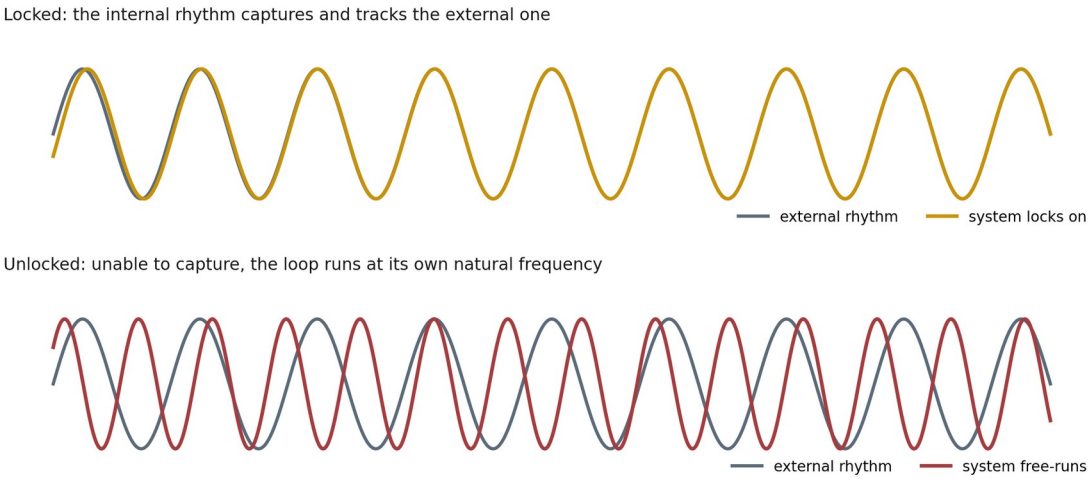


*Figure 4.2. Locking and free-running. When the loop can capture an external rhythm it tracks it. When it can't, it doesn't fall silent. It runs at its own natural frequency.*

Two cautions. Brains aren't phase-locked loops, and I'm using the term for a family of behaviors, capture range, lock range, free-running, rather than for a specific circuit. And the strongest supporting result cuts awkwardly. If autistic neural entrainment to rhythm really is reduced (Beker et al., 2021), reduced capture is supported, but the free-running account of stimming needs the internal oscillator to stay intact. Nobody has tested that.

The idea is worth chasing. Right now it's unsupported.

## 4.6a What this architecture commits us to

Stating the cost plainly. Once the prediction is the set point, the error the controller sees is a discrepancy between predicted and actual sensation rather than a gap between a state and an independently specified goal.

That's the active-inference architecture, and adopting it takes a side in a live dispute. Classical control engineering keeps the reference separate from the estimate, which is precisely what licenses tuning the estimator and the controller independently (Kalman, 1960; Joseph & Tou, 1961; Wonham, 1968). Friston's position is that biological systems have no such separate reference, because the goal is a prior inside the generative model rather than a cost outside it. Kennaway (2025) argues the two frameworks can't be reconciled at all.

The architecture drawn here is the biological one. Nothing hands an organism a reference signal. It infers what should be happening and acts to make the world match, and the sensation that tells it whether the action worked is the same sensation it used to form the belief.

What survives either way is the clinical distinction. Whether or not the two gains are formally separable in the mathematics, a clinician who can ask separately how confident is this belief and how hard is this person acting on it has two questions where the field has offered one.

## 4.7 What kind of claim this is

None of this is a new theory, and presenting it as one would be a mistake.

Every component is published. The estimation half is worked out in detail. The regulation half has a fifty-year literature behind it. The formal machinery for combining them turns sixty-six this year. Homeostatic set points have been treated formally as prior beliefs (Pezzulo, Rigoli & Friston, 2015; Stephan et al., 2016; Tschantz et al., 2022), and one paper states the division of labor in its title: "Interoception as modeling, allostasis as control" (Sennesh et al., 2022).

What hasn't been done is using the two-gain distinction as an organizing device for clinical thinking, and applying it to a phenotype where the controller-side signature has been measured four separate times and never named as such.

There's also a live objection that has to be met rather than dodged. Kennaway (2025) published a direct comparison of Perceptual Control Theory and the free energy principle and concluded they're fundamentally incompatible rather than complementary. Control theory sharply distinguishes action from prediction; the free energy framework treats action as a species of prediction. Friston's own position is stronger still. There's no separate controller to combine with an estimator, because the goal is a prior INSIDE the model rather than a cost outside it, which leaves precision as the only free parameter standing.

The position I'm taking is a third one. Whether the two halves are genuinely separable modules or one coupled process is an empirical question, not a matter of theoretical taste, and nobody has treated it as one.

The clinical value of the two-gain framing doesn't depend on winning that argument. Even if the brain doesn't implement two independent gains, a clinician who can ask separately how confident is this belief and how hard is this person acting on it has two questions where there used to be one.

Those two questions have different answers. They have different interventions. That's enough to be worth the trouble.

## 5. What this means for everyday practice: prediction, gain, and the maladaptive thought

What you have read so far isn't a treatment. It describes how a regulating system estimates its own state and then acts on the gap between that estimate and a set point. Nothing in it asks you to do anything you weren't already doing.

What it may do is tell you *why* the things you already do work, and why they fail in one specific, predictable way.

That claim is modest on purpose. **I see this as a complement (not a replacement) to cognitive behavioral therapy.** CBT has decades of outcome data behind it. A control-theoretic account of its mechanism has none, and an account that contradicted that outcome data would be the account with the problem. What follows maps the two vocabularies onto each other closely enough to think with, and marks at every point which parts are established, which are reasonable extension, and which are speculation.

The overlap you're about to notice is the point. If a mechanical model of regulation lands on the same interventions clinicians arrived at empirically over five decades, that's convergence, and convergence is the best evidence a young framework can offer.

---

### 5.1 The model I teach, and the same model drawn as a loop

Before the chain, the diagram. Here's what I've put in front of students and clients for years.

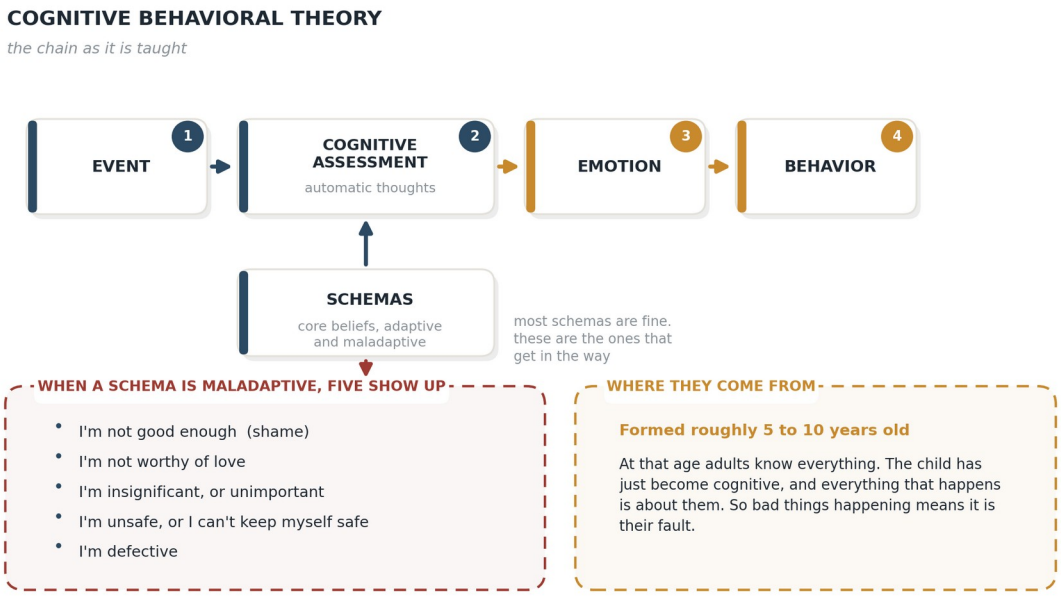


*Figure 5.1. The cognitive behavioral model as I teach it.*

An event happens. You run a cognitive assessment of it, which shows up as automatic thoughts. That assessment produces an emotion. The emotion produces behavior. Underneath the assessment sit the schemas, the core beliefs, and those are what the assessment is running on. When they're maladaptive, five show up over and over:

**I'm not good enough.** That one arrives as shame. **I'm not worthy of love. I'm insignificant, or unimportant. I'm unsafe, or I can't keep myself safe. I'm defective.**

They form early, roughly five to ten years old. From the child's point of view, at that age adults know everything. The child has just become cognitive enough to build general rules out of specific events, and not yet cognitive enough to consider that a rule might be about somebody else. So everything that happens is about them. Bad things happening means something is wrong with them, and the rule gets written down and filed. To be clear, not all core beliefs are maladaptive, it's the maladaptive ones that often get in our way.

Now redraw it.

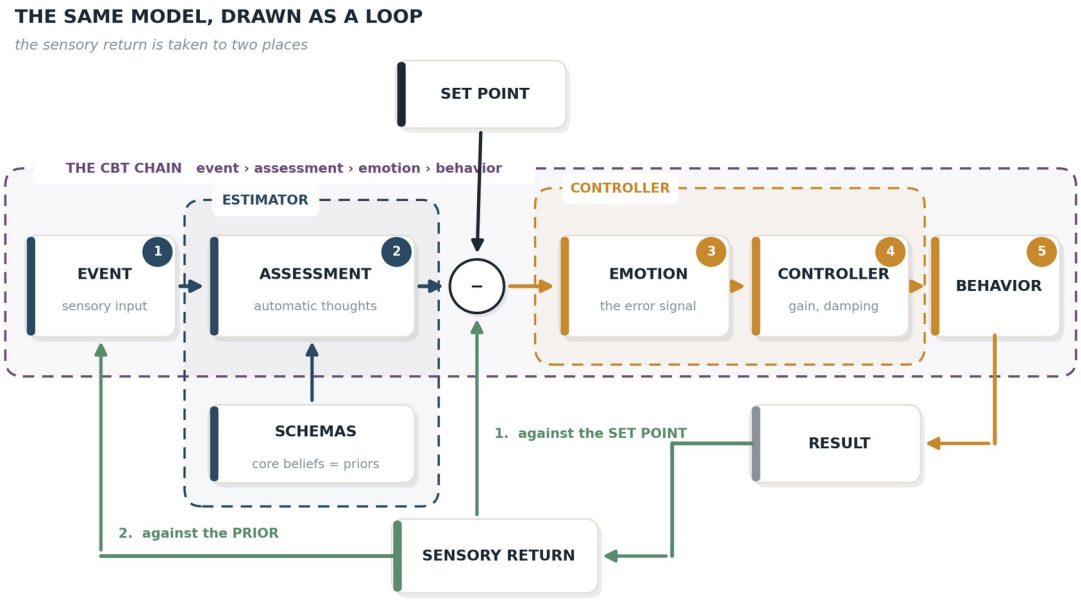


*Figure 5.2. The same model as a control loop.*

Every term carries over with nothing left standing. **The event** is sensory input. **The cognitive assessment** is the estimator, and **the schemas are the priors it runs on** — the same job, not a rename. **The emotion** is the error signal, the gap between the estimate and the set point being defended. **The behavior** is the controller's output.

Then the piece the standard diagram doesn't have. The behavior produces a **result**, and the result is the next event. The chain isn't a chain. It's a loop, and it runs continuously.

Two things follow.

**The step from assessment to emotion stops being a mystery.** In the standard teaching that arrow is the part that gets waved at, and students accept it because the sequence

obviously works even when the mechanism is vague. Under the control reading the emotion is a comparison, and its magnitude is the size of a discrepancy. Which is why the same thought produces a different feeling in two people, and a different feeling in one person on two different days. The thought didn't change. What it was being measured against did.

That also gives four of the five core beliefs a job and separates out the fifth. *Not good enough, not worthy of love, insignificant, defective* are priors, biasing what the estimator concludes from thin evidence. But *I'm unsafe, or I can't keep myself safe* isn't doing that. It sets how much estimated danger is tolerable before the system calls it an error. That's a set point, not a prior, and it needs a different intervention. Under the standard model all five sit in the same box.

**And the loop explains why a wrong belief survives contact with reality.** A chain ends at behavior, so reality never gets to answer.

Work an example. The prior is *I'm not worthy of love.* The event is a partner who's quiet at dinner. The estimator, running on that prior, reads quiet as withdrawal and withdrawal as the front edge of leaving. The set point is being securely connected. The gap is large, so the emotion is large, somewhere between fear and shame. The controller has high gain and no damping, so the reaction is to press. What's wrong. Are we okay. You'd tell me, right. Three hours of it. The result is a partner who was tired at the start of the evening and is now irritated and genuinely pulling back.

That result is the next event.

Nobody in that loop did anything irrational. Every step follows correctly from the step before it. The prior was wrong when the evening started and accurate by the time it ended, because the loop went and made it accurate.

That's what a maladaptive core belief is, mechanically. Not a wrong idea sitting in storage waiting to be corrected. A prior running a loop that manufactures its own evidence. Which is also why arguing with the belief so often does nothing: you're disputing a conclusion the system re-derives from fresh data every week.

---

## 5.2 Does the assessment do both jobs?

Here's the question I keep coming back to, and I don't have an answer to it.

The clean version of Figure 5.2 gives the cognitive assessment one job, which is to produce an estimate. But look at what an assessment actually delivers in a session. It tells you what's happening and how much it matters, in the same breath. *He's pulling away* is an estimate. *He's pulling away and this is the end of everything* is an estimate with a gain setting attached to it.

Appraisal research has said something close to this for decades. The appraisal is held to determine the intensity of the emotion and not only its content (Smith & Ellsworth, 1985).

If that's right, the assessment box is doing two mechanically different jobs, and the standard model is treating them as one.

Which opens something the diagram doesn't show. A maladaptive core belief may be able to enter the loop in three different places.

**As a prior.** *I'm defective* biases what the estimator concludes from thin evidence. Ambiguous feedback gets read as criticism.

**As a set point.** *I'm unsafe, or I can't keep myself safe* changes what counts as an error at all. Less estimated danger is tolerable, so the same situation opens a gap that wouldn't open in somebody else.

**As a gain.** *I'm not worthy of love* may not distort the reading whatsoever. It may set how hard the system acts once an error exists, because it reads the stakes as total. The estimate is accurate. The response is maximum.

Same five beliefs. Three mechanisms. Three different treatments, which is the argument of this entire document arriving from a direction I hadn't planned on.

I can't settle it from here and as far as I can tell nobody has. What's worth saying is that it's separable in principle. A belief operating as a prior predicts biased interpretation of *ambiguous* evidence and roughly normal handling of unambiguous evidence. A belief operating as a gain predicts accurate interpretation across the board and an outsized response to both. Hand the same person ambiguous and unambiguous versions of the same situation, and the two accounts come apart.

That's a study, and it's a small one.

**Status: Greg's hypothesis, untested.** Nobody has separated a belief's role as prior, set point and gain empirically. The design above would do it.

---

## 5.3 The chain, developed carefully

The clinical chain the model proposes runs like this:

> learning history → priors → predictions → the estimate the system regulates against → error signal → corrective action → the action prevents disconfirmation → the prior is reinforced

The arrows don't carry equal evidential weight, so take each one separately.

### Link 1: Learning history shapes priors

The oldest claim in cognitive therapy. Early and repeated experience builds durable representations that organize how a person reads everything after. Beck's schema construct says exactly that, and the generic cognitive model Beck and Haigh (2014) set out four decades later still says it, adding a modes-and-energizing account of why schemas activate in some contexts and stay dormant in others. Predictive processing calls those

same representations *priors*: stored expectations about what is likely, built from what has already happened. Renaming them changes nothing.

**Status: established.**

### Link 2: Priors generate predictions

Here the two vocabularies come apart in a way that matters. In the standard cognitive model, a schema is a stored belief that becomes available and biases appraisal once something triggers it. In the predictive account, the prior isn't sitting there waiting to be consulted. It is continuously *emitting* an expected sensory and interoceptive state, and perception is what you get when that expectation is reconciled with incoming evidence.

The practical difference is timing. Classic reading: the event happens, appraisal follows, feeling follows appraisal. Predictive reading: the expectation was already in place before the event, and the "event" the person experiences is the reconciled product.

Clients report the second version constantly. They tell you they walked into the room already braced. Appraisal-after-the-fact struggles with that. Prediction-before-the-fact doesn't.

**Status: established as a general account of perception; plausible extension as an account of clinical appraisal.** Its application to schema activation in a therapy room is an extension, not a finding.

### Link 3: Predictions set the estimate the system regulates against

This is the load-bearing move. A controller never acts on the world. It acts on the *estimate of the world* the observer hands it. If the estimate is off, the correction is off, however well the controller is tuned. Every engineer who has chased a control problem knows the failure often isn't in the controller at all. It's upstream, in what the controller was told.

Applied clinically: the anxious person isn't over-reacting to their situation. They're reacting accurately to an estimate of their situation that happens to be wrong.

DON'T say: "You're being irrational."
SAY: "I think your instrument is reading off."

That difference isn't cosmetic. Not conceptually, and not in the room.

**Status: the model's central proposal, and a reframe rather than a discovery.** Consistent with the cognitive model; not independently demonstrated as a description of clinical anxiety.

### Link 4: High gain on the resulting error drives intense corrective action

Two separable quantities enter here. **Observer gain** is how much new evidence is allowed to move the estimate, the precision term in predictive processing. **Controller gain** is how hard the system acts once an error is registered. A person can hold a wildly inaccurate

estimate and respond calmly, or hold an accurate estimate and respond violently, and you cannot tell which from the surface behavior. Section 5.5 develops the clinical consequences.

**Status: the observer/controller distinction is standard in control engineering and a plausible extension to affect regulation. Its clinical dissociation in humans has not, to the author's knowledge, been directly demonstrated.**

### Link 5: The corrective action prevents disconfirmation

This is the best-evidenced part of the chain, and control theory had nothing to do with it. It came from clinical observation of safety behaviors (Salkovskis, 1991) and it was tested experimentally: patients with panic disorder and agoraphobia asked to drop their safety-seeking behaviors during exposure showed greater reduction in catastrophic belief and anxiety than those who kept them (Salkovskis, Clark, Hackmann, Wells, & Gelder, 1999). In control language, the safety behavior removes the error signal without ever testing the estimate. The loop settles. Nothing is learned.

**Status: established, with an important qualification.** Rachman, Radomsky and Shafran (2008) argued that blanket rejection of safety behavior goes too far, and that judicious early use can help, partly because safety behavior doesn't always prevent disconfirmatory experience. State the mechanism properly and the argument stops being about whether safety behavior is bad: an action that terminates the error signal before the prediction is tested prevents updating. That's the whole rule.

### Link 6: The prior is reinforced

If the feared outcome doesn't occur *and* the person attributes its non-occurrence to the corrective action, the prior survives intact and gains a supporting instance. **Status: standard maintenance account in the cognitive model; needs no control-theoretic gloss to be true.**

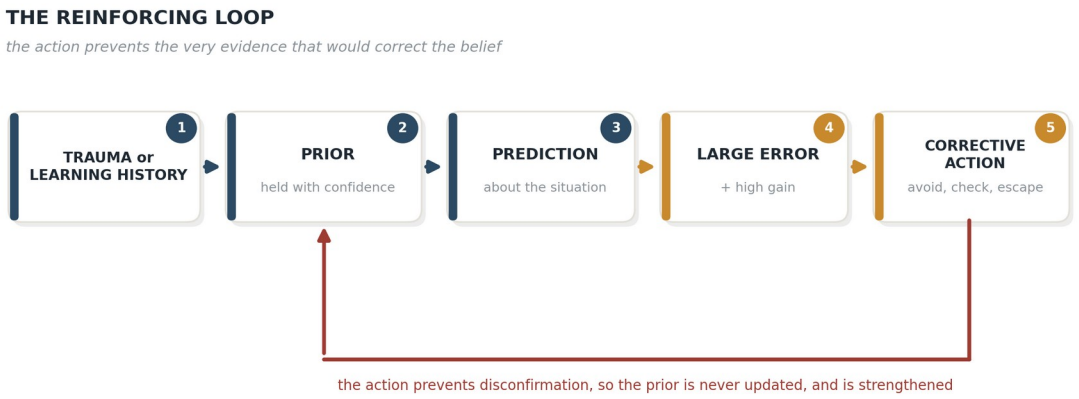


*Figure 5.3. The reinforcing loop. The corrective action succeeds in reducing distress, which is exactly why the belief that made it necessary is never tested.*

### Worked example: social anxiety (illustrative, not a real case)

An illustrative composite. An adult with a history of being publicly corrected at school carries a prior of the form *when I speak, people will judge what I said as stupid.* That prior doesn't get consulted after the meeting. It is emitted before the meeting begins, and it sets the expected state: hostile audience, imminent evaluation.

Because the prior is held with high confidence, the actual room barely moves the estimate. Neutral faces read as neutral-tending-hostile. A colleague looking at their phone gets scored as disengagement.

The error between "I am being judged" and "I need not to be judged" is large. Controller gain is high, so the correction is intense. Rehearse every sentence silently. Speak as little as possible. Keep it short. Monitor the self instead of the room.

The meeting ends without incident, and the prior survives. The person never tested *when I speak freely, people judge me*. They tested *when I speak in a heavily controlled way, nothing bad happens*, which is perfectly compatible with the prior. Self-monitoring also degraded the evidence stream that could have moved the estimate: they were watching themselves, not the audience, so the observer had almost no external data to work with.

Every element of that is available in the standard cognitive formulation. What the control reading adds is a statement about where the intervention has to land. Not on the content of the belief alone, but on the confidence it's held with and on the actions keeping it untested.

### Worked example: panic (illustrative, not a real case)

Clark's (1986) model of panic is catastrophic misinterpretation of benign bodily sensations. A palpitation is read as cardiac failure, the reading generates arousal, the arousal generates more sensation, the loop closes.

Read as a control problem, the front end of that loop is an interoceptive observer whose estimate is dominated by a prior of the form *this sensation means catastrophe*. Barrett and Simmons (2015) and Paulus and Stein (2006) both propose that interoception is substantially predictive rather than purely afferent, and that anxiety involves an altered interoceptive prediction signal. Then controller gain takes over: escape, seat-finding, breathing control, calling someone. The correction terminates the error, the catastrophe doesn't occur, the attribution goes to the correction, and the prior gains an instance.

**Be honest about what this is.** It restates Clark's model and predicts nothing Clark's model didn't. Its only potential advantage is making the *separateness* of two clinical targets explicit, confidence in the interoceptive prediction and intensity of the corrective response, where the original runs them together.

---

## 5.4 Where this agrees with what CBT already knows, and what it adds on top

### Schemas as priors: what the reframe actually adds

The lazy version says "schemas are priors" and stops. That's vocabulary substitution. Starbucks calling a small a tall doesn't put more coffee in the cup.

The addition is **precision**. In the standard cognitive model a belief has content and strength, and strength usually gets assessed on a 0–100 conviction rating. In the predictive account, the confidence attached to a prediction isn't a property of the content at all. It's a separate parameter determining how much weight the prediction gets against incoming evidence, and it can be moved without touching the content. The two dissociate constantly, and that dissociation explains one of the most common failures in a course of CBT.

Take two clients who both say "people think I'm stupid," both rating conviction at 70%. In one, the content is the problem: the belief is wrong, and contrary evidence gets in. Restructure the content and the estimate moves. In the other, the content may be partly accurate in some settings, but the prediction is held at such high precision that no evidence moves the estimate at all. Restructuring changes the words and leaves the parameter untouched. The client agrees with you and feels no different.

This also explains why conviction ratings are a poor guide to prognosis. I spent years reading a conviction rating as though it told me how hard a belief would be to shift. It doesn't. It tells you where the estimate currently sits. It says nothing about whether the estimate is MOVABLE, and movability is the thing you're actually trying to change.

**Status: the strongest thing the reframe offers, and still a conceptual claim rather than an empirical result.** No one has shown that a precision measure predicts treatment response better than a conviction rating. That study is doable. Nobody has done it.

### Safety behaviors and avoidance: the mechanism that prevents registration

Clients describe safety behaviors in the language of good sense, and the sentences are reasonable on their face. Here is the translation.

"I just wanted to be prepared." → I rehearsed until the prediction could no longer be tested.
"I only stayed twenty minutes, but I stayed." → I ended it before the error signal could arrive.
"I keep my phone in my hand, that's all." → I have an escape route, so nothing tonight counts as evidence.
"I sat near the door in case I needed air." → The outcome now belongs to the chair, not to the prediction.
"I didn't say much, and the meeting went fine." → I tested a different prediction from the one that's hurting me.

Every one of those removes the signal without testing the estimate.

The control account of safety behavior is direct. An error signal drives a corrective action. If that action removes the *signal* without testing the *estimate*, the loop reaches equilibrium

and no learning occurs. The system is working exactly as designed, regulating against a wrong reference. Salkovskis (1991) made the claim in cognitive terms and the 1999 experiment bore it out. It also clarifies why avoidance is more corrosive than distress: distress is uncomfortable but informative, while avoidance is comfortable, uninformative, and removes the data that would have updated the model.

Rachman et al. (2008) put a necessary limit on this. If the safety behavior still permits the disconfirming experience to register, it doesn't block learning, and it may enable the exposure that produces it. Stop auditing for the presence of safety behaviors. Audit for whether the prediction is being tested.

### The inhibitory learning model of exposure: the closest existing bridge

Here the control account and existing CBT theory aren't analogous. They're nearly identical, and that deserves care.

Craske, Treanor, Conway, Zbozinek and Vervliet (2014) set out an inhibitory learning approach to exposure, arguing that extinction doesn't erase the original fear association but creates a competing inhibitory one, so the therapeutic target is the *learning* of the new association and its later *retrieval*, not the reduction of within-session fear. Craske, Treanor, Zbozinek and Vervliet (2022) extended this into an inhibitory retrieval framework and the OptEx Nexus.

The first principle in the 2014 paper is **expectancy violation**. Exposure should maximize the mismatch between what the client expects and what actually happens, and should continue until that mismatch has been produced rather than until fear has subsided.

Read that in control terms. *Maximize the mismatch between expected and actual outcome* is prediction error maximization. Not a metaphor for it. The same quantity in a different formalism. Craske's model says learning is driven by the size of the mismatch. The predictive account says the estimate updates in proportion to precision-weighted prediction error. Same claim.

The clinical consequences follow just as directly from the control account, because every one of them exists to make sure a prediction error is generated and registered. Don't use fear reduction as the index of a successful exposure. Ask for an explicit prediction beforehand and check it afterwards. Vary the context. Combine feared stimuli. Remove safety signals. Occasionally reinforce rather than always disconfirm.

DON'T ask: "How anxious are you now compared with when we started?"
SAY: "What did you predict would happen? What actually happened?"

**This is the single most important paragraph in the section for a working clinician.** If the control model has clinical value, it's because it says the effective ingredient in exposure is error generation rather than habituation, and that claim already exists in the literature with evidence behind it. The control reading doesn't improve on Craske. It agrees with her.

## Behavioral experiments as prediction tests

The Oxford Guide (Bennett-Levy, Butler, Fennell, Hackmann, Mueller, & Westbrook, 2004) formalized behavioral experiments as planned activities designed to test the validity of a belief, with a prediction recorded in advance and an explicit review of the outcome against it. That's a prediction-error protocol written before the vocabulary existed. Record the prediction. Generate the observation. Compute the discrepancy. Update.

Reading it this way sharpens the account of why an experiment fails. It fails when the prediction was never stated precisely enough to be violated ("it will go badly" can't generate a clean error). It fails when precision on the prior is so high that the discrepancy gets absorbed rather than registered ("that was a fluke"). And it fails when a covert safety behavior meant the experiment tested a different prediction from the one written on the form.

All three are cheap to prevent, and the prevention is the same in every case: write the prediction down before anything happens. Skip that step and you get the Home Depot outcome. One trip for the part you thought you needed, two more for the parts the job actually required.

## Why insight alone often fails to change feeling

Clients say it routinely. "I know it's irrational but I still feel it." Stott (2007) called it rational–emotional dissociation. The predictive account gives a specific reading: verbal belief change can move a high-level prior, the one that generates sentences, and leave the precision of lower-level predictions exactly where it was. The explicit model has updated. The model actually running the interoceptive and affective estimate has not. Both are running, and only one of them talks.

Which lines up with something worth saying plainly. People decide emotionally and then build the logical case afterward. A Vulcan couldn't run a real business, because rationalization gets there ahead of logic almost every time. Every clinician reading this has watched a client assemble an airtight argument for a conclusion they had already reached before they sat down.

~~I am afraid to be wrong in front of these people.~~ "I just want to make sure we're all aligned before I commit."

The argument isn't the cause. It's the transcript. Treat the sentence as the output of the loop, and go after the loop.

This is also consistent with the dismantling evidence. Ougrin's (2011) meta-analysis of 20 randomized trials (n = 1,308) comparing cognitive therapy with exposure found no significant difference in panic disorder, PTSD or OCD, with cognitive therapy superior only in social phobia. If verbal restructuring were the active ingredient throughout, exposure shouldn't keep matching it. If error generation were, this is roughly the pattern to expect, with social phobia as the informative exception, because there the feared outcome is often ambiguous and hard to disconfirm behaviorally.

**Status: speculative interpretation.** The dismantling data are real. Reading them as support for a precision account rather than for any of several alternatives is a choice.

### Interoceptive exposure in panic as a precision problem

Interoceptive exposure, deliberately inducing sensations through hyperventilation, spinning, straw breathing, stair climbing, is a standard component of panic treatment.

Reframed: the client feeling their heartbeat isn't the problem. The interoceptive estimate is dominated by a high-precision prior about what the heartbeat means, and sensory evidence has almost no authority to move it. Interoceptive exposure forces a large, repeated, unambiguous mismatch between *this sensation means I am dying* and *this sensation preceded nothing*. It goes after precision, not content. The client already knows verbally that palpitations aren't fatal, and knowing hasn't helped. Fifty disconfirming instances delivered at the level where the prediction actually lives is what moves it.

WebMD runs the same architecture on purpose. Feed it one symptom and it hands back the worst diagnosis compatible with the input, every time, with total confidence. The difference is you can close the tab.

It also predicts a failure mode. Interoceptive exposure done with a concealed safety behavior, gripping the chair or breathing carefully or keeping a phone in hand, should produce less benefit, because the outcome gets attributed to the correction rather than to the falsity of the prediction. That's standard clinical teaching, and the control account gives the same answer.

---

## 5.5 The two dials, clinically

Separating observer gain from controller gain produces four intervention families instead of two, and makes visible a distinction that presents identically in the room.

**A client with an accurate belief and excessive controller gain needs a different intervention from a client with an inaccurate belief held at excessive precision.** Both arrive distressed. Both describe an intense reaction. Both may rate conviction at 80%. Restructuring is close to useless for the first and central for the second. Arousal work is central for the first and a distraction for the second.

The person whose workplace really is hostile, and who answers a real slight with three days of rumination and a resignation letter, doesn't have a distorted estimate. They have a correct estimate and a controller that overshoots. Telling them their thought is unbalanced is wrong and costly to the alliance. The target is the size of the corrective response and the damping on it. Reducing controller gain doesn't mean the response stops. Keith Richards is 82 with arthritis and books a residency instead of a world tour. The output continues. The amplitude comes down.

The person whose workplace is benign, who is certain it's hostile, and who has never let contradicting evidence in, has an estimate problem. Arousal regulation will make them more comfortable without touching the model producing the distress.

| Target | What it changes | Interventions | Evidence status |
|---|---|---|---|
| **Prior / content** | The expectation being emitted | Cognitive restructuring, psychoeducation, thought records, narrative and meaning-making work, formulation | Well-evidenced within CBT packages; not clearly separable from behavioral components in dismantling trials (Ougrin, 2011) |
| **Observer gain / precision** | The confidence the prediction is held with, independent of content | Behavioral experiments (Bennett-Levy et al., 2004); exposure designed for expectancy violation (Craske et al., 2014, 2022); attention training and mindfulness as precision allocation (Feldman & Friston, 2010); tolerating uncertainty (Carleton, 2016) | Exposure and behavioral experiments well-evidenced; the *precision* reading of their mechanism is a reframe, not a finding |
| **Controller gain** | Intensity of the corrective response, without touching the belief | Arousal regulation, paced breathing, applied relaxation, DBT distress tolerance skills, medication in some presentations | Component-level evidence varies; "reduces controller gain" is interpretation |
| **Damping** | Whether corrections converge or oscillate | Delay ("wait ten minutes"), urge surfing, response prevention, scheduled worry time, ritual postponement | Response prevention within ERP well-evidenced for OCD (Öst et al., 2015); "damping" is a reframe of the mechanism |

Three cautions about the table.

**The rows aren't clean.** A behavioral experiment changes precision and usually content too. Mindfulness changes attention allocation and also reduces arousal. The rows ask what an intervention is primarily for, not whether it does one job only.

**Breathing retraining sits awkwardly.** In panic it can function as a safety behavior, a corrective action that removes the error signal and blocks the test. Salkovskis, Clark and Hackmann (1991) reported that cognitive therapy for panic worked without exposure or breathing retraining, which should give pause to anyone reaching for it as a default. Reducing controller gain is useful when the belief is accurate and counterproductive when it becomes the thing preventing disconfirmation.

**The assessment question the table implies isn't one clinicians routinely ask.** Standard assessment asks what the client believes and how strongly.

DON'T stop at: "What do you believe, and how convinced are you?"
ALSO ASK: "What would it take to change your mind about that?" (precision). "What did you do about it?" (controller gain). "How many times did you do it?" (damping).

All three are answerable inside a clinical interview, and they point at different treatments.

---

## 5.6 Rumination, worry and compulsions as oscillation

A control loop with high gain and inadequate damping doesn't settle. It overshoots, corrects, overshoots the other way, and keeps going. That isn't a figure of speech borrowed for this chapter. It's the first thing you learn about feedback control, it's why damping gets designed into flight control systems deliberately rather than hoped for, and it's the most satisfying part of this whole analogy. Which is exactly why it needs handling carefully.

Rumination, worry and compulsive checking share a structure: a repetitive corrective action that never terminates because the error signal is never satisfied. The checker returns to the door because the state estimate, *the door is locked*, never reaches sufficient confidence. The action is performed, evidence is acquired, the estimate fails to stabilize, and the action repeats.

Fradkin, Adams, Parr, Roiser and Huppert (2020) built a computational account of OCD along these lines. Symptoms are modeled as arising from excessive uncertainty about *state transitions*, whether an action has actually changed the world, rather than from a distorted belief about the world's contents. If you don't trust that turning the key changed the lock's state, no amount of looking resolves the doubt, because looking is itself an action whose effect on your knowledge you also don't fully trust. The loop can't converge, because what is uncertain is the loop's own effect.

That's a far more precise claim than "OCD patients are anxious about germs," and it maps onto the damping question. The magnitude of the error isn't the problem. The corrective action doesn't reliably reduce it.

Watkins and Roberts (2020) review rumination as a transdiagnostic vulnerability with several maintaining mechanisms: habit formation, executive control, abstract processing style, goal discrepancies, negative bias. Their account is richer than "poor damping" and should be the primary source. The control reading adds only this much, that goal-discrepancy-driven repetitive thought has the shape of an unstable loop, an error signal the corrective action can't close, because thinking about a discrepancy doesn't resolve it.

### Why "not acting" is therapeutic in control terms

Response prevention is the part of ERP that clients hate and clinicians sometimes soft-pedal. The control reading gives it a clean rationale. Adding damping to an oscillating loop means reducing how quickly and forcefully the system acts on error, and doing nothing is the maximal case. Two things follow. The oscillation stops, because the repeated action was sustaining it. And the prediction finally gets tested, because the error signal persists long enough to be checked against reality instead of against the ritual.

The outcome data support the practice, though not this explanation of it. Öst, Havnen, Hansen and Kvale (2015) meta-analyzed 37 randomized trials of CBT for OCD published 1993–2014 and found very large effect sizes against waiting list (1.31) and placebo (1.33), significant superiority over antidepressant medication (0.55), and no significant difference between exposure and response prevention and cognitive therapy (0.07) or between individual and group delivery (0.17). The authors flag methodological problems across the trial set.

That last comparison should temper any strong claim. If ERP and cognitive therapy for OCD produce equivalent outcomes, a model explaining ERP's mechanism hasn't thereby explained OCD treatment. Whatever is doing the work is reachable through both routes.

The same reasoning underwrites urge surfing, worry postponement and the instruction to wait before acting. Each one inserts delay between error detection and corrective action, which is damping in control terms. Whether that's why they help isn't established.

---

## 5.7 Explaining the model TO clients

Three explanations follow that a therapist can say aloud. They're written as speech rather than prose, because speech is what has to survive contact with a room.

### The thermostat

> "A thermostat has two jobs. Work out what the temperature actually is, then decide how hard to run the heating. If the sensor reads five degrees cold, the room ends up too hot — and the heating isn't broken. It's doing exactly what it should, with bad information.
>
> Most of what we're doing here is checking two separate things. Is your reading accurate? And when it says there's a problem, how hard does your system respond? Different questions, different answers."

**What it does well:** it separates estimate from response without jargon, and it takes blame off the response. The heating isn't malfunctioning. That's often the sentence a client needs.

**Where it breaks:** thermostats don't learn, and the entire point of the clinical model is that this system learns and can be got to learn again. Use it for the two-part structure, then drop it before the client asks whether they're a machine.

### The smoke alarm

> "You've got a smoke alarm that goes off when you make toast. It is not a stupid alarm. It was set that sensitive for a reason — at some point, in your house, there really were fires. The setting was correct then.
>
> It hasn't been reset since, and every time it goes off you run out of the house before you can see whether there was a fire. So it never gets the information that would let it recalibrate. It's not that you're overreacting. It's that the alarm has never had the chance to find out."

**What it does well:** it locates the origin of the belief in real experience, which is usually true and usually a relief, and it makes the avoidance point without accusing anyone of avoiding.

**Where it breaks, and this one matters.** The framing implies the danger isn't real, which is exactly wrong for clients whose situations *are* dangerous. NEVER use it with someone in an ongoing abusive relationship, an unsafe workplace, or a community where their fear is well calibrated. For those clients the alarm is correct and the work is elsewhere. Get this wrong and you have told a person that their accurate perception of danger is a malfunction.

It also implies a single global sensitivity setting, when the clinical reality is domain-specific: exquisitely sensitive at work, fine at home. If the client raises that, take it seriously instead of defending the analogy.

### Why avoiding prevents updating

> "Every time you leave the meeting early, nothing bad happens — and your mind files it as *nothing bad happened because I left early*. That's a reasonable conclusion from the evidence you gave it. You just never ran the test you needed.
>
> The test is: stay, and find out. Not because staying will feel better — it probably won't, the first several times. It's the only version of the experiment that can produce new information. Right now your prediction has never once been allowed to be wrong."

**What it does well:** it reframes exposure as evidence gathering rather than endurance. The endurance framing sets up "how long until the anxiety drops?" as the measure of success, and that's precisely the index the inhibitory learning model says to abandon (Craske et al., 2014).

### On telling clients their belief is not stupidity

Worth saying explicitly, and early. A maladaptive belief in this model isn't a cognitive error or a failure of reasoning. It's a well-fitted model of experiences that actually happened, held

at a confidence level appropriate to those experiences, applied to a present that no longer matches.

Clients hear our vocabulary more accurately than we like. Here is what lands.

"That's a cognitive distortion." → You think badly.
"That's an unhelpful thinking style." → You think badly, politely.
"Let's challenge that thought." → I'm about to argue with you and I intend to win.
"Your model was accurate, and it's now over-confident." → What happened to you was real, and the only open question is the calibration.

DON'T say: "That's a cognitive distortion."
SAY: "Your model was accurate. It's over-confident now."

The second one is kinder and, under this account, closer to true. It makes the origin of the belief legitimate while leaving its current calibration open to question. You can respect where a belief came from and still test it.

### The trap

A highly verbal client will take the model and turn it into one more thing to think about. "I think my observer gain is high today" is not therapy. It's rumination with better vocabulary, a corrective action that terminates the error signal without testing anything.

Three guards. Don't introduce the model to a client whose presenting problem is over-analysis. Don't spend more than a few minutes on it, because it's scaffolding for a behavioral experiment and not content in its own right. And if model-language starts showing up in session, say out loud that describing the loop isn't the same as running the test.

The model earns its place if it makes a client more willing to run an experiment. If it makes them more interested in the explanation than in the experiment, it has cost you something.

---

## 5.8 Autistic clients specifically

Several of the adaptations below are already standard practice. What the predictive account contributes is a reason why they're needed, not a new technique.

### The evidence, reported honestly

Adapted CBT for anxiety in autistic people has a real evidence base, and it isn't uniform.

Sharma, Hucker, Matthews, Grohmann and Laws (2021) meta-analyzed 19 randomized controlled trials (833 participants; CBT n = 487, controls n = 346) of CBT for anxiety in autistic children and young people. Effect sizes diverged sharply by informant: large for clinician-rated symptoms (g = 0.88, 95% CI 0.55–1.12, k = 11), smaller but significant for parent-report (g = 0.40, 95% CI 0.24–0.56, k = 18), smaller again for child self-report (g = 0.25, 95% CI 0.06–0.43, k = 13). Benefits were **not maintained at follow-up**. Moderators

favored younger children and individual delivery. Using Cochrane Risk of Bias 2, the authors found concerns about reporting bias across most trials.

Perihan et al. (2020) meta-analyzed 23 studies of CBT for anxiety in children with what the paper terms high-functioning ASD: moderate pooled effect (g = −0.66), larger with parental involvement (g = −0.85) than child-only (g = −0.34), larger for standard-length (g = −1.02) than short-term (g = −0.37) interventions.

For adults, Weston, Hodgekins and Langdon (2016) reviewed 48 studies. For affective disorders: self-report g = 0.24 (non-significant), informant-report g = 0.66, clinician-report g = 0.73. For autism symptoms: self-report g = 0.25 (non-significant), informant-report g = 0.48, clinician-report g = 0.65, task-based g = 0.35. Sensitivity analyses reduced most of these, and the authors concluded that definitive trials are still needed.

Russell et al. (2013) randomized 46 adolescents and adults with autism and comorbid OCD to CBT or a matched anxiety-management control. Both worked: within-group effect sizes 1.01 for CBT and 0.60 for anxiety management, more responders in the CBT arm (45% versus 20%), but **no statistically significant between-group difference**. Self-rated improvement was small in both arms (0.33 and −0.05).

**Read that pattern carefully.** Across four syntheses, the effect is consistently largest when a clinician rates it and smallest, often non-significant, when the autistic person rates it themselves. That could mean autistic clients underestimate their gains, or that observers overestimate them, or that the measures miss what changed. Anyone who tells you which is guessing. The honest summary: adapted CBT for anxiety in autistic people has moderate support at end of treatment, weak evidence of maintenance, and a systematic discrepancy by informant.

For mindfulness, Hartley, Dorstyn and Due (2019) pooled 10 studies (233 autistic children and adults, 241 caregivers) and found significant post-intervention gains in subjective wellbeing, apparently maintained at three months, while noting that more controlled research is needed. Promising rather than settled.

### Literalness

Every analogy in 6.5 is a liability. A client who processes the thermostat or the smoke alarm literally may leave believing you said their brain is a machine, or that their fear is false.

Adjustments. Say the thing without the metaphor first, then offer the metaphor as an optional aid and check what landed. Avoid stacked analogies. Be exact about degree: “this will probably feel uncomfortable for the first ten to fifteen minutes” is usable, “it’ll get easier” is not. Write the prediction and the outcome down, because a written record is a better error signal than a remembered impression, and it removes the demand to reconstruct an emotional state from memory.

### Interoceptive differences and alexithymia as a confound

This adaptation is substantial, because it undercuts a core CBT procedure.

Kinnaird, Stewart and Tchanturia (2019) meta-analyzed alexithymia in autism using the Toronto Alexithymia Scale across 15 studies and found roughly 50% prevalence in autistic samples (49.93%) against 4.89% in neurotypical comparison groups. Common, not universal, and on their reading a subgroup with specific clinical needs rather than a defining feature of autism.

The interoception picture is less clear than it's commonly asserted to be. Klein, Witthöft and Jungmann (2025) reviewed 31 studies and meta-analyzed the five adult studies with comparable measures, finding **no significant difference** in cardiac interoceptive accuracy between autistic adults and neurotypical controls ($\hat{\mu} = -0.21$, SE = 0.11, p = .06). Anyone claiming autistic interoception is straightforwardly impaired is ahead of the data.

The clinical implication doesn't depend on resolving that. Standard CBT asks the client to identify an emotion, rate its intensity, locate the bodily sensations, and link them to a thought. For a client with marked alexithymia that isn't a therapeutic task. It's an assessment of a capacity they may not have, handed over as homework. Failure then gets misread as resistance or as lack of insight.

Workarounds. Shift from emotion labels to observable behavior and situation, so "what did you do next?" rather than "what did you feel?" Use physiological or contextual anchors instead of introspective ones. Accept a coarse binary, tolerable or not tolerable, instead of a 0–100 scale. Allow written or asynchronous reporting. And take seriously that a client may know their behavior changed without being able to name the feeling. That is usable clinical data. Don't spend six sessions converting it into an emotion word.

### The risk of treating a sensory-regulatory need as a maladaptive behavior

This is the most serious avoidable harm in the section.

A safety behavior, in the model, is a corrective action that removes an error signal without permitting a test. Repetitive movement, ear defenders, leaving a loud room, a fixed routine: these superficially resemble avoidance, and a clinician working from an anxiety protocol will be tempted to target them. Some of them aren't avoidance of a feared outcome at all. They're direct regulation of a genuinely aversive sensory environment. The distinction is the one drawn in 6.3. **Is the estimate wrong, or is it accurate and the response proportionate to a real input?** If a fluorescent-lit open-plan office is actually painful for this person, leaving it is a correct response to a correct estimate, and applying response prevention asks someone to endure pain in order to disconfirm a prediction that happens to be true.

The test before targeting any behavior: *what prediction is this action preventing from being tested, and is that prediction false?* If you can't name a false prediction, you aren't looking at a safety behavior. Leave it alone.

### Why environmental predictability may do more than restructuring

Van de Cruys et al. (2014) proposed that autism involves inflexibly high precision on prediction errors, errors treated as more informative than they should be, so the system keeps trying to explain noise. Under that account an unpredictable environment isn't

merely unpleasant. It's a continuous stream of high-weighted errors demanding resolution. Reducing that unpredictability through stable routines, advance notice of change, explicit agendas, consistent sensory conditions, and clear rules rather than inferred social ones cuts the error load at source, where restructuring adjusts the interpretation of a stream still arriving at full volume.

**This is speculation and should be labeled as such in any clinical conversation.** The mechanistic argument is coherent. The comparative claim, that predictability work outperforms restructuring for autistic clients, has never been tested head to head. What can be said with more confidence is that the practitioners surveyed by Spain et al. (2023) in a three-round Delphi study identified a range of accommodations needed to make CBT acceptable and effective for autistic clients. Expert consensus, not trial evidence.

A defensible position: treat environmental predictability as a first-line target alongside standard CBT components rather than instead of them, and don't assume a client who improves when their environment stabilizes has had a cognitive change.

---

## 5.9 What this does NOT license

The preceding pages are a way of thinking. They aren't a warrant for anything, and the distance between those two things needs stating plainly.

**No outcome trial has tested this model.** Every effect size reported here belongs to an existing, independently developed treatment: CBT, exposure, ERP, adapted CBT for autistic clients. None of them was derived from a control-theoretic account, and none of them validates one. The evidence supports the treatments and says nothing at all about the explanation offered here for why they work.

**Nobody has shown that explaining this model to clients improves outcomes.** The language in 6.5 exists because clinicians asked for something sayable. It has never been compared against another explanation, or against giving none. It might help. It might waste session time. It might hand the wrong client a framework for intellectualizing. There's no data either way, and a clinician using it is running an uncontrolled experiment on their own caseload.

**It doesn't justify new techniques.** If a technique follows from the model but has no independent evidence, the model is not evidence for it. This is where mechanistic reframes historically go wrong. The account feels explanatory, a procedure gets derived from it, and the coherence of the account gets mistaken for support for the procedure. Every intervention in the 6.3 table has its own evidence base or lacks one, and the model changes neither.

**Do not present it to clients as established neuroscience.** Predictive processing is a live and contested research framework, not settled fact, and precision-weighting in clinical anxiety is an inference rather than a measurement. Saying "your brain is doing Bayesian inference" asserts more than anyone knows, and it makes the therapist's authority the

reason for believing it. Use it as a way of thinking, "one way to look at this is…", and drop it the moment it stops helping.

**It doesn't replace formulation.** A control diagram contains no history, no relationships, no meaning, and no account of what the person wants. Treat it as the formulation and you strip out most of what makes a formulation useful.

### What would need to be true for this to earn clinical status

Four things, in rough order of tractability:

1. **A measure of precision separable from conviction.** How movable a belief is, independent of where it currently sits, with evidence that it behaves differently from a 0–100 conviction rating.
2. **Prospective prediction.** That measure, taken at baseline, should predict who responds to belief-focused versus exposure-focused work. If it doesn't predict differential response, the distinction isn't clinically real however elegant it looks.
3. **Dissociation of the two gains in humans.** Evidence that observer gain and controller gain can be independently manipulated and measured in a clinical sample. Without this, the two-dial framing is a heuristic and not a mechanism.
4. **A treatment-selection trial.** Clients allocated by gain profile versus allocated as usual, outcomes compared. This is the only test that would make the model clinically consequential rather than merely coherent.

Until at least the first two exist, the correct claim is narrow. This organizes things clinicians already do, it may sharpen the choice between them, and it may be wrong.

---

## 6. The model outside the clinic: decisions, teams and organizations

This part surprised me. The same four parameters describe a person deciding under pressure, a manager running a team, and a whole company steering itself. Not by analogy either. Same mechanics, because they're all regulators with sensors, targets and lag.

### The lag problem, which is most of it

Go back to the person with the thermostat dial. They aren't stupid, and they aren't over-emotional. They're acting on a system whose response arrives later than their expectation of it, and they don't know that yet.

That is the single most common failure I see in businesses.

A marketing change goes in on Monday. Results land in four to six weeks. The owner checks on Wednesday, sees nothing, and changes it again. Then again the following week. By the time the first change would have shown up, three more have been layered on top and no result can be attributed to anything. The campaign didn't fail. The loop was run faster than the system's response time.

The fix isn't patience as a virtue, it's damping as an engineering decision. **Find out what the system's response time actually is, then set your review interval longer than that.** A change reviewed on a shorter cycle than its own feedback delay can't be evaluated. It can only be replaced.

### The two dials at work

The separation between estimate and response is at least as useful in an office as in a session, because the two failures look identical from the outside and get the same unhelpful advice.

**An accurate read with too much gain.** A leader who correctly notices a real problem and responds at a magnitude the problem doesn't warrant. The estimate is right. Telling them they're overreacting to nothing is both wrong and expensive, because they aren't reacting to nothing. What needs to come down is the amplitude, not the perception. Reducing controller gain doesn't mean the response stops. Keith Richards is 82 with arthritis and books a residency instead of a world tour. The output continues. The amplitude comes down.

**An inaccurate read held with high confidence.** Someone certain about a market, a competitor or a person, on thin evidence, who doesn't update when contradicted. Arousal work does nothing here. What moves it is a test with a written-down prediction.

Same complaint from a colleague — "he overreacts" — two different repairs. Ask which one you're looking at before you intervene.

## Micromanagement is a gain and damping problem

A manager who checks work every two hours is running a control loop at a frequency far above the rate at which the work actually produces signal. Every check generates an error, every error generates a correction, and the corrections arrive faster than the work can absorb them. The employee's output starts oscillating, which the manager reads as unreliability, which raises the checking rate.

The loop is the problem, not either person. Lengthen the interval and the oscillation damps out on its own. That's hard to do because it means acting less at the exact moment the system looks worst… which is the same thing that makes response prevention hard for a client.

## KPIs are set points, and most of them are wrong

An organization steering by numbers is a control system with all the same failure modes.

**The set point is arbitrary.** A number picked to be motivating rather than to describe a defended state. The system will defend it anyway. **The sensor measures a proxy.** Plasma osmolality is not thirst, and monthly revenue is not business health. Both get defended as though they were the thing. **The delay is unacknowledged.** Reporting lag means every correction is aimed at where the business was, not where it is. **The gain is too high.** One bad month triggers a response sized for a bad year. **There's no damping.** Strategy changes every quarter, and no strategy runs long enough to produce evidence about itself.

A company with those five settings looks chaotic, and it usually gets described in terms of culture and personality. It's a loop with the parameters set wrong. Parameters you can adjust. Culture you mostly can't.

## The reinforcing loop, in a business

The clinical loop has an exact commercial twin, and it's worth stating because it's expensive.

A belief that a market segment won't buy produces a prediction, which produces reduced effort with that segment, which produces poor results, which confirms the belief. The action prevented the evidence. Nobody was irrational at any point, and the belief is now backed by data the belief itself generated.

The test is the same as it is in a session: state the prediction in advance, run the experiment without the safety behavior, and check the result against what you wrote down rather than against how it felt.

## Decisions under pressure

Three things fall out that are worth carrying into a hard decision.

**Separate what you think is happening from how hard you're going to act on it.** Say them as two sentences. Most bad decisions under pressure are an accurate read with a disproportionate response, and they get defended by pointing at the accuracy of the read.

**Ask what your review interval is before you start.** If you don't set it in advance you'll set it under stress, and under stress it will be too short.

**Ask what would change your mind, and write it down before you have the answer.** That converts a belief into a test. It's the only reliable way to tell whether you're updating on evidence or defending a prior.

None of that is new advice. What the model adds is why it works, and that matters at the moment you least want to follow it.

# 7. Conclusions and applications

## 7.1 The claim, on one page

A biological control system regulates variables it can't measure directly. It has to estimate them, and the estimate is a weighted blend of what it expected and what it sensed.

Four parameters govern what happens next.

**Observer gain** sets how much the senses move the estimate relative to expectation. **Controller gain** sets how forcefully the system acts on the gap between that estimate and its set point. **Damping** decides whether corrections settle or overshoot. **Delay** decides how stale the information driving each correction already is.

All four are learned, from the statistics of the environment the person actually lived in. Volatility teaches observer gain. Consequence teaches controller gain. Overshoot teaches damping. Demand moves the set point.

Emotion is the error signal in that system, computed on an inference rather than a measurement. Which is why emotion responds to reframing at all, and why it so reliably fails to respond enough.

The autism application proposes that some autistic presentations involve an over-gained, under-damped loop, with repetitive behavior as the oscillatory signature and sensory-perceptual differences sitting on the observer side.

The clinical application proposes that a learning history sets a prior, the prior generates a prediction, the prediction determines the error, high gain drives an intense corrective action, and the action prevents the disconfirmation that would have revised the prior.

## 7.2 Four questions this hands you

The practical yield isn't a technique. It's a set of questions that separate presentations which look identical from across the room.

**Is this belief inaccurate, or is it accurate and being acted on too hard?** A client whose read of a situation is broadly correct but whose response runs at four times the required force needs something different from a client whose read is distorted. Cognitive restructuring addresses the second and does very little for the first. The first one needs work on the intensity of the response. Arousal regulation, distress tolerance, deliberate delay.

**Is the belief wrong, or is it held with too much confidence?** Content and confidence come apart. A moderately accurate belief held with total certainty behaves worse than a somewhat inaccurate belief held loosely, because certainty blocks updating. Behavioral experiments work on confidence rather than content. Which is exactly why they outperform argument.

**Is the person failing to update, or being prevented from updating?** Avoidance and safety behavior stop the error signal from ever being generated. The prior isn't stubborn. It's starved.

**What environment taught this setting, and does that environment still exist?** Almost every maladaptive parameter was adaptive somewhere. Naming the environment that taught the setting is usually more useful than disputing the setting, and it's a great deal less adversarial.

## 7.3 Where it applies

**In psychotherapy generally.** The model gives you a mechanism story for why exposure and behavioral experiments work. They generate prediction error, and prediction error is the only thing that revises a prior. It also tells you why insight so often doesn't. And it gives you a non-pejorative account of maladaptive belief as an over-confident model built from real evidence, which most clients accept far more readily than the suggestion that their thinking is distorted.

**With autistic clients.** The most actionable implication is that predictability and sensory intensity are DIFFERENT VARIABLES and should be intervened on separately. A quiet room that's unpredictable can be worse than a noisy one that isn't. It also supplies a mechanistic reason for caution about suppressing stimming. If the behavior is a regulatory output, taking it away without addressing the error it's correcting leaves the error uncorrected and removes the correction.

That deserves more than a caution, because most clinicians were trained to treat stimming as a symptom to reduce and were never told what it does. The companion autism document carries the full first-person record. The short version is that autistic adults describe at least five distinct jobs: regulating arousal in BOTH directions rather than only calming, setting a rhythm the body organizes around, filtering the environment so attention can hold, heading off escalation while the error is still small, and signaling to people who know them that the load is climbing. One participant in Kapp et al. (2019) described the rhythm function better than the literature does, saying a stim "sort of metronomes everything in your body to sort of go at that speed." Another found the early-correction function by accident: it "actually managed to help me stave off some panic attacks."

Some stims injure, and that's a real clinical problem. The model's answer is specific. **If a stim is causing harm, the target is the harm, not the regulation.** Identify what the loop is correcting and supply another output that does the same job. Substitute, don't subtract.

**In formulation.** The parameter set works as a scaffold. Which variable is being regulated, toward what set point, using what evidence, held with how much confidence, corrected how hard, and what's preventing the correction from ever being tested. That's a more mechanical formulation than most models hand you, and it points at different intervention targets depending on where the answer sits.

**In psychoeducation.** The thermostat is a good teaching object because everybody already owns one. The extension, a thermostat whose thermometer is guessing, takes about a minute to explain and tends to stick.

### 7.4 The limits

This is a mechanistic reframe. It isn't a treatment, it has no outcome data, and no trial has tested whether explaining it to a client improves anything at all. Use it as a way of thinking about a case. Don't present it to a client as established neuroscience, and don't let it license new techniques.

---

## 8. Where this all falls apart

Any framework worth using should be examined for how it fails. This one has at least eight ways, and several of them are serious.

### 8.1 The flexibility problem

A model with free parameters for set point, observer gain, controller gain, damping, delay, and arbitration threshold can be fitted to almost any behavior you like. Run into an unexpected result and there's always a parameter to adjust. That's the objection Bowers and Davis (2012) raised against Bayesian modeling generally, that priors, likelihoods and utility functions can be tuned to accommodate any dataset. Adding a control layer makes it worse, not better, because it adds parameters.

The only defense is committing in advance to which parameter explains which phenomenon, and specifying what result would rule the account out. Easy to state. Hard to hold. Judge this document on whether its predictions are genuinely falsifiable or only look that way.

### 8.2 The identifiability problem doesn't go away

In a Gaussian update only the RATIO of sensory precision to prior precision enters the posterior. A flatter prior and a sharper likelihood produce identical predictions. That has bedeviled the autism literature for fourteen years and it applies here unchanged. For most data, "they trusted their senses too much" and "they held their expectation too loosely" are the same claim wearing different clothes. Separating them needs paradigms that elicit calibrated confidence alongside estimates, and those almost never get run.

### 8.3 The autism evidence is reinterpretation, not prediction

The four findings supporting an over-gained under-damped loop were collected to answer other questions and interpreted in other terms. Reinterpreting existing data is legitimate, and it's how fields often move. It's also weaker than predicting a result before you see it,

and it's wide open to selection. Four findings that fit aren't impressive if there are forty that don't and nobody counted.

Nobody counted.

### 8.4 The separation may not be real

Friston's argument is that the brain has no cost function outside the generative model, and therefore no separate controller to bolt an estimator onto. Baltieri and Buckley (2020) show what that costs formally: active inference ties the control weight to sensory precision and computes neither gain explicitly. Kennaway (2025) argues from the other direction that the frameworks are fundamentally incompatible.

If either one is right, the central move in this document is an engineering convenience projected onto a system that doesn't work that way. The clinical questions in 7.2 would survive it. The mechanistic story behind them would not.

### 8.5 It has almost nothing to say about meaning

Control systems regulate variables. A great deal of what happens in a consulting room is narrative, relationship, identity, moral injury, grief, and the interpretation of a life. Set points and gains don't reach any of that, and a clinician who adopted this vocabulary as a complete account of the work would be poorer for it.

This describes a regulatory layer. It isn't a theory of persons.

### 8.6 The autism application risks re-pathologizing

A model describing autistic regulation as an under-damped loop sits one careless step from describing autistic people as badly tuned. The learning-theoretic framing helps, since parameters are correct solutions to prior environments rather than defects. It doesn't eliminate the risk.

Todorova and colleagues (2024) consulted autistic people about a closely related theory and found it captured part of the experience while missing how motivated autistic people are to explore and engage. Any framework built on the assumption that a difference is a mis-setting inherits the question of who decides what the correct setting is. Crippen (2026) argues that predictive-processing accounts stipulate that threshold rather than deriving it, and then read the diagnosis back out of it.

### 8.7 Several load-bearing sources are weak

*The Mind in the Wheel* is a self-published blog serial with no peer review and essentially no new data. Two of the central papers connecting active inference to classical control are still arXiv preprints. The stimming literature is overwhelmingly qualitative. The sensory integration evidence base shows study quality and positive findings running in opposite directions.

A synthesis is only as strong as what goes into it, and several of these inputs are not strong.

### 8.8 Nobody has tested any of it

Not the oscillation account of repetitive behavior. Not the phase-locking account of social interaction. Not the observer/controller dissociation. Not the clinical chain.

This is a framework proposal. Read it that way the whole way through.

---

## 9. Where do we go from here

### 9.1 Six studies

**Frequency-domain reanalysis of existing motor data.** The force-tracking and postural datasets already collected contain what's needed to tell noise-tracking instability from resonant instability. A resonant peak at a characteristic frequency implicates controller gain and damping. Broadband variance scaling with sensor noise implicates observer gain. No new participants required. This could be run out of existing archives.

**Stereotypy as a dependent variable in gain manipulation.** The visual-display-gain paradigm already exists and already shows autistic force irregularity worsening as displayed error gets amplified. Nobody has measured repetitive behavior during those sessions. Add that one measure and you test the model's central prediction directly.

**A damping intervention.** Introduce lag, smooth the feedback, drop its resolution, then measure stability and repetitive behavior. This model predicts improvement. Every account treating the problem as excess sensory information predicts the opposite. That makes it genuinely discriminating.

**The four-condition stimming experiment.** Compare naturally chosen stimming against matched non-repetitive movement, against externally imposed rhythmic stimulation, and against stillness, while manipulating environmental predictability. The externally-imposed-rhythm condition is the critical one, because it separates rhythmicity from self-generation and simultaneously pits the phase-locking account against the predictability account. They make opposite predictions about it.

**Confidence calibration.** Any paradigm eliciting calibrated confidence alongside perceptual estimates separates over-precise sensory evidence from attenuated priors. That's been the available discriminating test for over a decade and it mostly hasn't been run.

**A clinical dissociation study.** If observer gain and controller gain separate in real people, belief accuracy and response intensity should dissociate across clients, and the two subgroups should respond differently to cognitive versus arousal-focused intervention. That's a testable prediction about treatment matching, and it's the only claim in here that would change practice if it held.

## 9.2 What would have to be true first

Four things, in order.

The parameters would have to be measurable in individual people with acceptable reliability. Measured differences would have to predict something clinically meaningful that existing measures miss. Intervening on a specific parameter would have to beat treatment as usual. And the whole apparatus would have to survive the discovery that some of its components don't separate.

None of that has happened.

Until it does, this is a way of thinking about a case. Not a tool for treating one.

## 9.3 What's worth doing now

Three things don't require waiting on any of the above.

Ask the two questions separately. How confident is this belief, and how hard is this person acting on it. The answers differ, the interventions differ, and asking costs you nothing.

Treat predictability and intensity as distinct variables when you're working with sensory distress. Intervene on them separately instead of assuming less stimulation is always the answer.

And when a regulatory pattern looks maladaptive, ask what environment would have made it correct. That question is usually answerable, usually informative, and it changes the conversation from correction to recognition.

## 9.4 A closing note

Here's where I'm supposed to hand you the framework that ties it together and tell you it changes everything.

I'm not going to.

This document has argued that a framework's capacity to absorb findings can outrun its capacity to predict them, and that a theory which grows after every disconfirmation is being protected rather than tested. Those criticisms apply to what I've proposed here at least as much as to the literature it reviews.

The right response to an elegant model isn't adoption. It's asking what the thing forbids.

This one forbids repetitive behavior with no frequency structure. It forbids damping interventions failing while sensory reduction succeeds. It forbids observer and controller differences that never dissociate. It forbids gain manipulations leaving repetitive behavior untouched.

If those results come in, the model is wrong and should be thrown out rather than patched.

That's the standard I'd hold it to. It's also the standard the existing literature in this area has largely failed to meet, which is most of why I wrote any of this down.

## Appendix A. How this differs from *Psycho-Cybernetics*

Anyone who has spent time in the self-help section will have noticed that a plastic surgeon got to this vocabulary in 1960 and sold more than thirty million copies with it. Maxwell Maltz's *Psycho-Cybernetics* borrowed Wiener's engineering language, applied it to the self, and became one of the most influential business and sports psychology books ever written. It deserves a straight answer rather than a sneer, so here it is.

**What Maltz got right.** His central claim is that behavior organizes itself around a defended reference, and that the reference sets the ceiling rather than effort does. That's a set point, described in a mass-market paperback a decade before control theory reached clinical psychology in any serious way. He also saw that the system is goal-seeking and self-correcting rather than driven, which is the same insight Powers formalized in 1973. The observation was sound. Coaches have been getting mileage out of it for sixty years for a reason.

**What's missing is the estimator.** Maltz has a servo-mechanism and a target and essentially nothing between the world and the target. There is no separate machinery that takes sensory evidence, weighs it against a prior, and produces an estimate that the set point then tracks. Which is precisely the hole this document exists to fill. Without an estimator you cannot distinguish a person whose reading of the situation is wrong from a person whose reading is right and whose response is too large, and that distinction is the entire clinical payload of Section 5. Maltz collapses both into self-image.

Three further differences follow from that one.

**One master set point instead of many.** Maltz runs everything through a single global reference, the self-image, and treats it as the thing that determines outcomes across domains. The account here has many parallel governors, each defending its own variable, competing for one body. That's why a person can be professionally confident and socially terrified at the same time without any contradiction, which a single-self-image model has to explain away.

**Set points move on imagination.** Maltz's mechanism of change is mental rehearsal: picture the outcome vividly enough and the servo will steer toward it. Under this account the estimate moves on evidence weighted by precision, which is why visualizing yourself calm at the edge of a roof does close to nothing and standing at the edge of a roof does a great deal. The distinction isn't academic. It is the difference between a technique with fifty years of outcome trials behind it and a technique without.

**There is no evidence base.** Maltz worked from case observation, uncontrolled, with no falsification criteria and no way to be wrong. Section 8 of this document lists four ways the model proposed here could be shown false. *Psycho-Cybernetics* offers none, and a system that cannot fail a test is not making a claim about the world.

## The twenty-one days

Maltz is also the origin of the most repeated false number in behavior change, and the story is worth telling accurately because it is the same failure this whole document keeps finding in the research literature.

What Maltz actually wrote is that it requires a minimum of about twenty-one days for an old mental image to dissolve and a new one to jell. He was describing how long his surgical patients took to stop seeing their old face in the mirror, and how long amputees took to stop feeling a limb that was gone. A careful observation about psychological adjustment to a changed body, hedged twice, in the same sentence.

What got repeated dropped “minimum,” dropped “about,” and swapped “mental image” for “habit.” A clinical field note became a law of behavior change, and it is now printed on wall calendars.

The actual figure comes from Lally, van Jaarsveld, Potts and Wardle (2010), who tracked 96 people adopting a new daily behavior and measured the time to automaticity. The median was 66 days. The range ran from 18 to 254. There is no single number, which is the finding, and any program built on a fixed one is built on a misquotation of a surgeon’s observation about faces.

Maltz was more careful than the people who cite him. That pattern recurs throughout this document, and on nearly every page of the companion literature review. The damage is almost never done by the person who made the original claim. It is worth noticing that the most durable damage in a field usually isn’t done by the person who made the original claim.

## References

*108 sources, verified against Crossref, Europe PMC, PubMed or the publisher record.*


Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry, 4*, 47. https://doi.org/10.3389/fpsyt.2013.00047

Angeletos Chrysaitis, N., & Seriès, P. (2023). 10 years of Bayesian theories of autism: A comprehensive review. *Neuroscience & Biobehavioral Reviews, 145*, 105022. https://doi.org/10.1016/j.neubiorev.2022.105022

Ashby, W. R. (1952). *Design for a brain*. Chapman & Hall. https://doi.org/10.5962/bhl.title.6969

Baltieri, M., & Buckley, C. L. (2020). *On Kalman-Bucy filters, linear quadratic control and active inference* (arXiv:2005.06269) [Preprint]. arXiv. https://arxiv.org/abs/2005.06269

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience, 12*(11), 1833. https://doi.org/10.1093/scan/nsx060

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience, 16*(7), 419–429. https://doi.org/10.1038/nrn3950

Barrett, L. F., Quigley, K. S., & Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1708), 20160011. https://doi.org/10.1098/rstb.2016.0011

Beck, A. T., & Haigh, E. A. P. (2014). Advances in cognitive theory and therapy: The generic cognitive model. *Annual Review of Clinical Psychology, 10*, 1–24. https://doi.org/10.1146/annurev-clinpsy-032813-153734

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience, 10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Beker, S., Foxe, J. J., & Molholm, S. (2021). Oscillatory entrainment mechanisms and anticipatory predictive processes in children with autism spectrum disorder. *Journal of Neurophysiology, 126*(5), 1783–1798. https://doi.org/10.1152/jn.00329.2021

Bennett-Levy, J., Butler, G., Fennell, M., Hackmann, A., Mueller, M., & Westbrook, D. (Eds.). (2004). *Oxford guide to behavioural experiments in cognitive therapy*. Oxford University Press. https://doi.org/10.1093/med:psych/9780198529163.001.0001

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138*(3), 389–414. https://doi.org/10.1037/a0026450

Brown, H., Adams, R. A., Pareés, I., Edwards, M., & Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing, 14*(4), 411–427. https://doi.org/10.1007/s10339-013-0571-3

Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience, 18*(4), 590–596. https://doi.org/10.1038/nn.3961

Büchel, C., Geuter, S., Sprenger, C., & Eippert, F. (2014). Placebo analgesia: A predictive coding perspective. *Neuron, 81*(6), 1223–1239. https://doi.org/10.1016/j.neuron.2014.02.042

Cannon, J., O'Brien, A. M., Bungert, L., & Sinha, P. (2021). Prediction in autism spectrum disorder: A systematic review of empirical evidence. *Autism Research, 14*(4), 604–630. https://doi.org/10.1002/aur.2482

Cannon, W. B. (1929). Organization for physiological homeostasis. *Physiological Reviews, 9*(3), 399–431. https://doi.org/10.1152/physrev.1929.9.3.399

Carleton, R. N. (2016). Into the unknown: A review and synthesis of contemporary models involving uncertainty. *Journal of Anxiety Disorders, 39*, 30–43. https://doi.org/10.1016/j.janxdis.2016.02.007

Carver, C. S., & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality–social, clinical, and health psychology. *Psychological Bulletin, 92*(1), 111–135. https://doi.org/10.1037/0033-2909.92.1.111

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clark, D. M. (1986). A cognitive approach to panic. *Behaviour Research and Therapy, 24*(4), 461–470. https://doi.org/10.1016/0005-7967(86)90011-2

Colombo, M., & Seriès, P. (2012). Bayes in the brain — On Bayesian modelling in neuroscience. *The British Journal for the Philosophy of Science, 63*(3), 697–723. https://doi.org/10.1093/bjps/axr043

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers, A. R. (2019). Hallucinations and strong priors. *Trends in Cognitive Sciences, 23*(2), 114–127. https://doi.org/10.1016/j.tics.2018.12.001

Craik, K. J. W. (1943). *The nature of explanation*. Cambridge University Press.

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014). Maximizing exposure therapy: An inhibitory learning approach. *Behaviour Research and Therapy, 58*, 10–23. https://doi.org/10.1016/j.brat.2014.04.006

Craske, M. G., Treanor, M., Zbozinek, T. D., & Vervliet, B. (2022). Optimizing exposure therapy with an inhibitory retrieval approach and the OptEx Nexus. *Behaviour Research and Therapy, 152*, 104069. https://doi.org/10.1016/j.brat.2022.104069

Crippen, M. (2026). Errors or adaptations? A critical review of predictive processing in psychiatry. *Behavioral Sciences, 16*(7), 1116. https://doi.org/10.3390/bs16071116

Crompton, C. J., Ropar, D., Evans-Williams, C. V., Flynn, E. G., & Fletcher-Watson, S. (2020). Autistic peer-to-peer information transfer is highly effective. *Autism, 24*(7), 1704–1712. https://doi.org/10.1177/1362361320919286

Cui, K., Lin, X., Gao, R., Jing, S., Luo, F., & Wang, J. (2026). A systematic review and meta-analysis of empirical evidence for the simple Bayesian model of autism. *Neuropsychology Review, 36*(2), 184–195. https://doi.org/10.1007/s11065-025-09672-8

Edwards, M. J., Adams, R. A., Brown, H., Pareés, I., & Friston, K. J. (2012). A Bayesian account of 'hysteria'. *Brain, 135*(11), 3495–3512. https://doi.org/10.1093/brain/aws129

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4*, 215. https://doi.org/10.3389/fnhum.2010.00215

Feldman, R. (2012). Parent–infant synchrony: A biobehavioral model of mutual influences in the formation of affiliative bonds. *Monographs of the Society for Research in Child Development, 77*(2), 42–51. https://doi.org/10.1111/j.1540-5834.2011.00660.x

Fradkin, I., Adams, R. A., Parr, T., Roiser, J. P., & Huppert, J. D. (2020). Searching for an anchor in an unpredictable world: A computational model of obsessive compulsive disorder. *Psychological Review, 127*(5), 672–699. https://doi.org/10.1037/rev0000188

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology, 3*, 130. https://doi.org/10.3389/fpsyg.2012.00130

Furutachi, S., & Hofer, S. B. (2026). Rethinking predictive processing. *Annual Review of Neuroscience, 49*(1), 471–494. https://doi.org/10.1146/annurev-neuro-102124-031410

Gizowski, C., & Bourque, C. W. (2018). The neural basis of homeostatic and anticipatory thirst. *Nature Reviews Nephrology, 14*(1), 11–25. https://doi.org/10.1038/nrneph.2017.149

Hartley, M., Dorstyn, D., & Due, C. (2019). Mindfulness for children and adults with autism spectrum disorder and their caregivers: A meta-analysis. *Journal of Autism and Developmental Disorders, 49*(10), 4306–4319. https://doi.org/10.1007/s10803-019-04145-3

Higginson, S., Mansell, W., & Wood, A. M. (2011). An integrative mechanistic account of psychological distress, therapeutic change and recovery: The Perceptual Control Theory

approach. *Clinical Psychology Review, 31*(2), 249–259. https://doi.org/10.1016/j.cpr.2010.01.005

Hohwy, J. (2016). The self-evidencing brain. *Noûs, 50*(2), 259–285. https://doi.org/10.1111/nous.12062

Joseph, P. D., & Tou, J. T. (1961). On linear control theory. *Transactions of the American Institute of Electrical Engineers, Part II: Applications and Industry, 80*(4), 193–196. https://doi.org/10.1109/TAI.1961.6371743

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, 82*(1), 35–45. https://doi.org/10.1115/1.3662552

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron, 100*(2), 424–435. https://doi.org/10.1016/j.neuron.2018.10.003

Kennaway, R. (2025). Perceptual control theory and the free energy principle: A comparison. *Current Opinion in Behavioral Sciences, 65*, 101575. https://doi.org/10.1016/j.cobeha.2025.101575

Kinnaird, E., Stewart, C., & Tchanturia, K. (2019). Investigating alexithymia in autism: A systematic review and meta-analysis. *European Psychiatry, 55*, 80–89. https://doi.org/10.1016/j.eurpsy.2018.09.004

Klein, M., Witthöft, M., & Jungmann, S. M. (2025). Interoception in individuals with autism spectrum disorder: A systematic literature review and meta-analysis. *Frontiers in Psychiatry, 16*, 1573263. https://doi.org/10.3389/fpsyt.2025.1573263

Lafer-Sousa, R., Hermann, K. L., & Conway, B. R. (2015). Striking individual differences in color perception uncovered by 'the dress' photograph. *Current Biology, 25*(13), R545–R546. https://doi.org/10.1016/j.cub.2015.04.053

Lally, P., van Jaarsveld, C. H. M., Potts, H. W. W., & Wardle, J. (2010). How are habits formed: Modelling habit formation in the real world. *European Journal of Social Psychology, 40*(6), 998–1009. https://doi.org/10.1002/ejsp.674

Lawson, R. P., Mathys, C., & Rees, G. (2017). Adults with autism overestimate the volatility of the sensory environment. *Nature Neuroscience, 20*(9), 1293–1299. https://doi.org/10.1038/nn.4615

Lidstone, D. E., Miah, F. Z., Poston, B., Beasley, J. F., Mostofsky, S. H., & Dufek, J. S. (2020). Children with autism spectrum disorder show impairments during dynamic versus static grip-force tracking. *Autism Research, 13*(12), 2177–2189. https://doi.org/10.1002/aur.2370

Litwin, P., & Miłkowski, M. (2020). Unification by fiat: Arrested development of predictive processing. *Cognitive Science, 44*(7), e12867. https://doi.org/10.1111/cogs.12867

Maltz, M. (1960). *Psycho-cybernetics*. Prentice-Hall.

Mansell, W., & Marken, R. S. (2015). The origins and future of control theory in psychology. *Review of General Psychology, 19*(4), 425–430. https://doi.org/10.1037/gpr0000057

Mansell, W., Gulrez, T., & Landman, M. (2025). The prediction illusion: Perceptual control mechanisms that fool the observer. *Current Opinion in Behavioral Sciences, 62*, 101488. https://doi.org/10.1016/j.cobeha.2025.101488

Marken, R. S. (2013). Taking purpose into account in experimental psychology: Testing for controlled variables. *Psychological Reports, 112*(1), 184–201. https://doi.org/10.2466/03.49.PR0.112.1.184-201

Marken, R. S. (2014). Testing for controlled variables: A model-based approach to determining the perceptual basis of behavior. *Attention, Perception, & Psychophysics, 76*(1), 255–263. https://doi.org/10.3758/s13414-013-0552-8

Marken, R. S., & Mansell, W. (2013). Perceptual control as a unifying concept in psychology. *Review of General Psychology, 17*(2), 190–195. https://doi.org/10.1037/a0032933

Maxwell, J. C. (1868). On governors. *Proceedings of the Royal Society of London, 16*, 270–283. https://doi.org/10.1098/rspl.1867.0055

Milton, D. E. M. (2012). On the ontological status of autism: The "double empathy problem." *Disability & Society, 27*(6), 883–887. https://doi.org/10.1080/09687599.2012.710008

Öst, L.-G., Havnen, A., Hansen, B., & Kvale, G. (2015). Cognitive behavioral treatments of obsessive-compulsive disorder: A systematic review and meta-analysis of studies published 1993–2014. *Clinical Psychology Review, 40*, 156–169. https://doi.org/10.1016/j.cpr.2015.06.003

Ougrin, D. (2011). Efficacy of exposure versus cognitive therapy in anxiety disorders: Systematic review and meta-analysis. *BMC Psychiatry, 11*, 200. https://doi.org/10.1186/1471-244X-11-200

Pareés, I., Brown, H., Nuruki, A., Adams, R. A., Davare, M., Bhatia, K. P., Friston, K., & Edwards, M. J. (2014). Loss of sensory attenuation in patients with functional (psychogenic) movement disorders. *Brain, 137*(11), 2916–2921. https://doi.org/10.1093/brain/awu237

Parr, T., & Friston, K. J. (2017). Uncertainty, epistemics and active inference. *Journal of the Royal Society Interface, 14*(136), 20170376. https://doi.org/10.1098/rsif.2017.0376

Paulus, M. P., & Stein, M. B. (2006). An insular view of anxiety. *Biological Psychiatry, 60*(4), 383–387. https://doi.org/10.1016/j.biopsych.2006.03.042

Perihan, C., Burke, M., Bowman-Perrott, L., Bicer, A., Gallup, J., Thompson, J., & Sallese, M. (2020). Effects of cognitive behavioral therapy for reducing anxiety in children with high functioning ASD: A systematic review and meta-analysis. *Journal of Autism and Developmental Disorders, 50*(6), 1958–1972. https://doi.org/10.1007/s10803-019-03949-7

Petzschner, F. H., Weber, L. A. E., Gard, T., & Stephan, K. E. (2017). Computational psychosomatics and computational psychiatry: Toward a joint framework for differential diagnosis. *Biological Psychiatry, 82*(6), 421–430. https://doi.org/10.1016/j.biopsych.2017.05.012

Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology, 134*, 17–35. https://doi.org/10.1016/j.pneurobio.2015.09.001

Plank, I. S., Pior, A., Yurova, A., Nowak, J., Shi, Z., & Falter-Wagner, C. M. (2026). Predicting emotional valence in autism: A preregistered study in the Bayesian Brain framework. *Molecular Autism, 17*(1), 32. https://doi.org/10.1186/s13229-026-00730-3

Popper, K. R. (2002). *The logic of scientific discovery*. Routledge. (Original work published 1935)

Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. *Science, 357*(6351), 596–600. https://doi.org/10.1126/science.aan3458

Powers, W. T. (1973a). *Behavior: The control of perception*. Aldine.

Powers, W. T. (1973b). Feedback: Beyond behaviorism. *Science, 179*(4071), 351–356. https://doi.org/10.1126/science.179.4071.351

Powers, W. T. (1978). Quantitative analysis of purposive systems: Some spadework at the foundations of scientific psychology. *Psychological Review, 85*(5), 417–435. https://doi.org/10.1037/0033-295X.85.5.417

Powers, W. T., Clark, R. K., & McFarland, R. L. (1960). A general feedback theory of human behavior: Part I. *Perceptual and Motor Skills, 11*(1), 71–88. https://doi.org/10.2466/pms.1960.11.1.71

Rachman, S., Radomsky, A. S., & Shafran, R. (2008). Safety behaviour: A reconsideration. *Behaviour Research and Therapy, 46*(2), 163–173. https://doi.org/10.1016/j.brat.2007.11.008

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87. https://doi.org/10.1038/4580

Rescorla, R. A. (1972). Informational variables in Pavlovian conditioning. *Psychology of Learning and Motivation, 6*, 1–46. https://doi.org/10.1016/S0079-7421(08)60383-7

Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science, 10*(1), 18–24. https://doi.org/10.1086/286788

Rozin, P. (1968). Are carbohydrate and protein intakes separately regulated? *Journal of Comparative and Physiological Psychology, 65*(1), 23–29. https://doi.org/10.1037/h0025404

Russell, A. J., Jassi, A., Fullana, M. A., Mack, H., Johnston, K., Heyman, I., Murphy, D. G., & Mataix-Cols, D. (2013). Cognitive behavior therapy for comorbid obsessive-compulsive disorder in high-functioning autism spectrum disorders: A randomized controlled trial. *Depression and Anxiety, 30*(8), 697–708. https://doi.org/10.1002/da.22053

Salkovskis, P. M. (1991). The importance of behaviour in the maintenance of anxiety and panic: A cognitive account. *Behavioural Psychotherapy, 19*(1), 6–19. https://doi.org/10.1017/S0141347300011472

Salkovskis, P. M., Clark, D. M., & Hackmann, A. (1991). Treatment of panic attacks using cognitive therapy without exposure or breathing retraining. *Behaviour Research and Therapy, 29*(2), 161–166. https://doi.org/10.1016/0005-7967(91)90044-4

Salkovskis, P. M., Clark, D. M., Hackmann, A., Wells, A., & Gelder, M. G. (1999). An experimental investigation of the role of safety-seeking behaviours in the maintenance of panic disorder with agoraphobia. *Behaviour Research and Therapy, 37*(6), 559–574. https://doi.org/10.1016/S0005-7967(98)00153-3

Sapey-Triomphe, L.-A., Bouet, R., Mattout, J., Sonié, S., Schmitz, C., & Lecaignard, F. (2025). Systematic review and meta-analysis of mismatch negativity in autism: Insights into predictive mechanisms. *Autism Research, 18*(12), 2431–2450. https://doi.org/10.1002/aur.70131

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schwartz, S., Shinn-Cunningham, B., & Tager-Flusberg, H. (2018). Meta-analysis and systematic review of the literature characterizing auditory mismatch negativity in individuals with autism. *Neuroscience & Biobehavioral Reviews, 87*, 106–117. https://doi.org/10.1016/j.neubiorev.2018.01.008

Sennesh, E., Theriault, J., Brooks, D., van de Meent, J.-W., Barrett, L. F., & Quigley, K. S. (2022). Interoception as modeling, allostasis as control. *Biological Psychology, 167*, 108242. https://doi.org/10.1016/j.biopsycho.2021.108242

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17*(11), 565–573. https://doi.org/10.1016/j.tics.2013.09.007

Seth, A. K., & Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1708), 20160007. https://doi.org/10.1098/rstb.2016.0007

Seth, A. K., Millidge, B., Buckley, C. L., & Tschantz, A. (2020). Curious inferences: Reply to Sun and Firestone on the dark room problem. *Trends in Cognitive Sciences, 24*(9), 681–683. https://doi.org/10.1016/j.tics.2020.05.011

Sharma, S., Hucker, A., Matthews, T., Grohmann, D., & Laws, K. R. (2021). Cognitive behavioural therapy for anxiety in children and young people on the autism spectrum: A systematic review and meta-analysis. *BMC Psychology, 9*(1), 151. https://doi.org/10.1186/s40359-021-00658-8

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology, 48*(4), 813–838. https://doi.org/10.1037/0022-3514.48.4.813

Spain, D., Milner, V., Mason, D., Iannelli, H., Attoe, C., Ampegama, R., Kenny, L., Saunders, A., Happé, F., & Marshall-Tate, K. (2023). Improving cognitive behaviour therapy for autistic individuals: A Delphi survey with practitioners. *Journal of Rational-Emotive & Cognitive-Behavior Therapy, 41*(1), 45–63. https://doi.org/10.1007/s10942-022-00452-4

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A. E., Paliwal, S., Gard, T., Tittgemeyer, M., Fleming, S. M., Haker, H., Seth, A. K., & Petzschner, F. H. (2016). Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience, 10*, 550. https://doi.org/10.3389/fnhum.2016.00550

Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology & Behavior, 106*(1), 5–15. https://doi.org/10.1016/j.physbeh.2011.06.004

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., & Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological Psychiatry, 84*(9), 634–643. https://doi.org/10.1016/j.biopsych.2018.05.015

Stott, R. (2007). When head and heart do not agree: A theoretical and clinical analysis of rational-emotional dissociation (RED) in cognitive therapy. *Journal of Cognitive Psychotherapy, 21*(1), 37–50. https://doi.org/10.1891/088983907780493313

Sun, Z., & Firestone, C. (2020). The dark room problem. *Trends in Cognitive Sciences, 24*(5), 346–348. https://doi.org/10.1016/j.tics.2020.02.006

Todorova, G. K., Hatton, R. E. M., Sadique, S., & Pollick, F. E. (2024). The world is nuanced but pixelated: Autistic individuals' perspective on HIPPEA. *Autism, 28*(2), 498–509. https://doi.org/10.1177/13623613231176714

Tschantz, A., Barca, L., Maisto, D., Buckley, C. L., Seth, A. K., & Pezzulo, G. (2022). Simulating homeostatic, allostatic and goal-directed forms of interoceptive control using active inference. *Biological Psychology, 169*, 108266. https://doi.org/10.1016/j.biopsycho.2022.108266

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review, 121*(4), 649–675. https://doi.org/10.1037/a0037665

Van de Cruys, S., Friston, K. J., & Clark, A. (2020). Controlled optimism: Reply to Sun and Firestone on the dark room problem. *Trends in Cognitive Sciences, 24*(9), 680–681. https://doi.org/10.1016/j.tics.2020.05.012

Watkins, E. R., & Roberts, H. (2020). Reflecting on rumination: Consequences, causes, mechanisms and treatment of rumination. *Behaviour Research and Therapy, 127*, 103573. https://doi.org/10.1016/j.brat.2020.103573

Weston, L., Hodgekins, J., & Langdon, P. E. (2016). Effectiveness of cognitive behavioural therapy with people who have autistic spectrum disorders: A systematic review and meta-analysis. *Clinical Psychology Review, 49*, 41–54. https://doi.org/10.1016/j.cpr.2016.08.001

Wiener, N. (1948). *Cybernetics: Or control and communication in the animal and the machine*. MIT Press.

Wonham, W. M. (1968). On the separation theorem of stochastic control. *SIAM Journal on Control, 6*(2), 312–326. https://doi.org/10.1137/0306023

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron, 46*(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026

Zimmerman, C. A., Leib, D. E., & Knight, Z. A. (2017). Neural circuits underlying thirst and fluid homeostasis. *Nature Reviews Neuroscience, 18*(8), 459–469. https://doi.org/10.1038/nrn.2017.71