Steering Inside Determinism

Control, Goals and Free Will in a Brain That Obeys Physics

Brad Caldwell
caldwbr@gmail.com

November 2026

Abstract

Every part of a brain obeys physical law. It is often concluded from this either that no one steers anything, or that steering needs a gap in physics through which a self can act. Both conclusions assume that steering takes a chooser. It does not. Feedback makes an outcome follow a state the system holds, and not the system's history, with no chooser anywhere in it. We work this out for a brain. A goal fixes a few dimensions of what happens and is held in the format of perception. Built-in needs have their own cells, which fire until relief is predicted. A chosen goal is a pattern of activity that a circuit keeps alive, and in the fruit fly the cells that hold the goal, hold the estimate and compare the two are all known. How a brain like ours checks a goal it has made up against the present is not known, and we say what we propose. Many goals of several kinds are in force together, and deciding is their interplay. What a person gains is self-determination with a definite range. It is not authorship of oneself from nothing. The paper sets out the control theory it uses, a classification of goals, a model in which plasticity trades range against control, and nine predictions for Ring Bank Theory.

1 A thermostat steers

A thermostat is built from parts that only obey physics. A strip of metal bends as it warms, the bend closes a switch, and the switch runs a furnace. Yet the room ends at the temperature on the dial whether the day is cold or mild, whether a window is open, and whatever temperature the room started at. Turn the dial and the room ends somewhere else.

So the device is fully determined and it steers. The two do not compete, because they answer different questions. Determinism is about what each part obeys. Steering is about what the outcome depends on. A feedback loop changes the second and leaves the first alone.

By *loop* we mean a circle of cause. The furnace changes the room, the thermometer reads the room, and the reading switches the furnace. It is a loop through the world and back, and not a ring of neurons, although a brain builds its loops out of neurons (Section 5).

1.1 The experiment

Figure 1 strips this down to one quantity, random pushes on it, and a set point, which is the engineer's word for the setting. Each line is one run, and the ten runs start at ten different values. With the loop off, each run ends where its pushes leave it, and the endings cover 0.95 of the range. With the loop on, the same starts and the same pushes end within 0.04 of the range, at the set point. New pushes leave them there, and moving the set point moves all ten.

The figure is also a test for what an outcome depends on: change one thing, hold the rest, and see whether the outcome moves. Changing the set point moves it. Changing the starts or the pushes

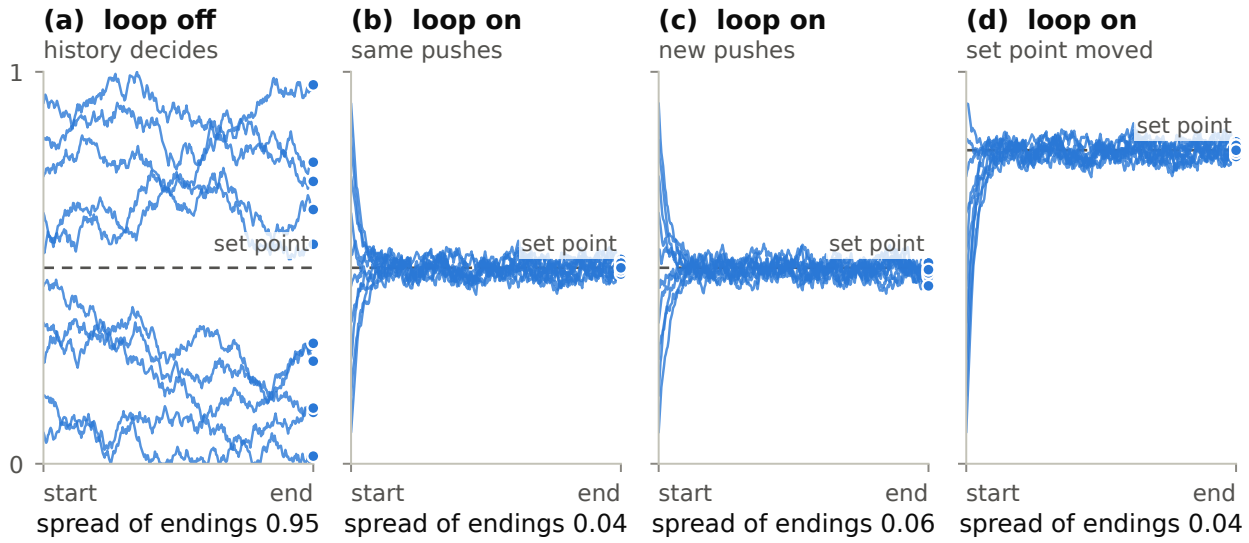

Figure 1: Ten runs of one simple system. Each line is a run, from its start on the left to its end on the right (dots). (a) Loop off: the endings follow the starting points and the pushes. (b) Loop on, with the same starts and the same pushes: the endings follow the set point (dashed). (c) New pushes: the endings stay. (d) The set point is moved and nothing else: the endings move with it. The model is one line (Appendix A).

does not. This is the interventionist test of causation, on which a cause is something one could change in order to change the effect [1]. We will call it the wiggle test.

> DEFINITION. **Steering.** A system *steers* an outcome when it holds a setting such that (i) changing the setting changes the outcome, and (ii) changing the starting point and the disturbances does not. The system is *self-determining* to the extent that the setting is a state of the system itself.

A thermostat steers and is not self-determining, since someone else sets its dial. A brain sets its own (Section 7).

1.2 The setting is physical, and it is still the right cause

The setting is nothing over and above the physics. It is where the dial points. Naming it as the cause adds no force to the world. It identifies the part of the physical state that the outcome follows.

A complete list of particle positions would also be a true answer to "why 70?" It would be a worse one, because nearly all of it could have been different with the same result. A cause should be no more specific than it needs to be [2]. When an outcome stays the same across every fine-grained way of realizing a state, the state is what makes the difference [3]. Measured by how much an intervention narrows down what happens next, a coarse description of such a system can carry more causal information than the fine one [4]. A reader who rejects the interventionist account of causation will not accept this step, and the paper depends on it.

1.3 What steering costs

A loop makes the ending forget the beginning. Runs that start apart and get the same input are pulled together exponentially [5]. Erasing differences costs energy: erasing one bit of information dissipates a minimum of heat [6], and the furnace bill is the same fact at scale.

1.4 Where the paper goes

The rest applies this to a brain.

Section 2 The worry that physics leaves no room for steering, and the ghost brought in to answer it, rest on the same false premise.

Section 3 The control theory used: gain, delay, prediction, stacked loops. A brain's signals are slow, so the "now" it acts in is an estimate.

Section 4 A brain's setting is a goal. It fixes a few things, leaves the rest open, and is held as what it would look and feel like to have it.

Section 5 How a brain holds one, in the cases where the cells are known. Needs have their own cells. A chosen goal is a pattern of activity that a circuit keeps alive. One step is still not understood.

Section 6 Kinds of goals: reached or kept, starting behavior or only restraining it, stored in habit or in a memory that lapses. Many are in force at once.

Section 7 Deciding is the interplay of the goals in force. A person's freedom has the range of the goals their brain can voice, and the range can be widened.

Section 8 What that amounts to as free will, and what it does not.

Section 9 The loop's own settings change with age, sleep and chemistry. Plasticity buys range and costs control.

Section 10 Consequences for Ring Bank Theory, and nine predictions.

How statements are marked. Established results are in the running text with citations. Four kinds of statement are boxed: a *Definition* fixes a word, an *Observation* is the author's own observation of experience, a *Proposal* is our reading and could be wrong, and a *Hypothesis* is a guess stated so that it can be tested. Statements about the brain are candidates, and the models are toys that show an effect exists and which way it goes.

2 The worry, and why a ghost seemed necessary

2.1 The worry

If every particle in a brain follows physical law, then what the brain does next is fixed by where its particles are now. So how can a person steer anything? We call this the *Newton worry*. It has two usual answers.

The passenger. The first answer gives up the steering. The brain does what its particles do, and experience is a display on which nothing depends. A person watches their life and steers none of it.

The ghost. The second answer keeps the steering and gives up on physics being the whole story. The one who steers is no part of the brain. In its plainest form the idea runs like this. I am one thing, present through the whole brain. Wherever a neuron is close to firing, so that almost nothing would tip it either way, I settle whether it fires, and through a great many such settlings I steer. Quantum mechanics comes into it because it is the one place where physics itself seems to leave an outcome open, so a settling there would break no law.

Two published proposals come close to this. Beck and Eccles proposed that a mental intention acts by briefly raising the probability of transmitter release at the synapses of a cortical cell, the release being modeled as a quantum event [7]. In the theory of Hameroff and Penrose, quantum

processes in the microtubules of neurons end in reductions of the quantum state, and these are taken to be moments of awareness or of choice [8]. Hameroff has argued that the selections can set off the firing of axons, that they may escape algorithmic determinism, and that the theory can in this way rescue conscious free will [9]. Ryle called the doctrine of a mind that stands apart from the body it operates "the dogma of the ghost in the machine" [10]. We will call the picture the *ghost*.

The ghost is attractive because it looks like what real steering would take: steering that is one's own, and that nothing fixed in advance.

2.2 One mistake, made twice

PROPOSAL. **The passenger and the ghost share a premise.** Both answers assume that steering takes a *chooser*: a party distinct from the mechanism who settles what the mechanism does. The passenger view finds no chooser and concludes there is no steering. The ghost view finds steering and concludes there is a chooser. The premise is false: the loop of Section 1 steers with no chooser in it.

View	Is there steering?	Is there a chooser?	"Steering takes a chooser"
the passenger	no	no	assumed
the ghost	yes	yes	assumed
the loop (this paper)	yes	no	denied

Table 1: Three answers to the Newton worry.

2.3 Why a ghost would not help

It would have to be a controller. To steer a brain toward anything, a ghost would have to want something, register how things stand, and push so as to close the gap. That is a setting, a sensor and a controller, which is a loop. Positing a ghost explains nothing that the brain's own loops leave unexplained.

It could act only through the settings. A single neuron is a small lever. The firing of a cortical neuron varies greatly from one occasion to the next, and quantities appear to be carried by ensembles of tens of neurons [11]. To change what a brain does, the ghost would have to change what the brain is trying to do. The brain's own loops change that already.

"Totally free" is a coin toss. The ghost is wanted because its choices are supposed to be fixed by nothing. Take that literally. A choice that depends on nothing does not depend on what the person wants, knows or cares about. That is a coin toss, and a coin toss steers nothing. In a loop, an event that nothing determines is a disturbance: something to be corrected. It is not a source of control. One reply is that the selections are not random, only not computable [9]. Another is that a choice left open between purposes that are both the agent's own is still the agent's [12]. We come back to that one in Section 8, because it is the serious version.

The physics is disputed, and the argument does not need it. One calculation puts the lifetimes of the relevant quantum states in a brain at 10^{-13} to 10^{-20} seconds, against a thousandth to a tenth of a second for the workings of neurons [13]. A reply holds that the calculation does not

address the model actually proposed, and that corrected times are far longer [14]. We take no side. Even if the quantum openings are there, the three points above stand.

2.4 Why it feels as if there is a chooser

Part of the answer is known. The feeling of having willed an act is itself an inference. People draw it when a thought comes just before an act, fits the act, and has no visible rival [15]. The machinery that raised the thought is out of view. Ring Bank Theory's name for the source of what one is about is "the invisible upstream" [16, §1.3]. A thought arrives, an act follows, and nothing else is in sight to credit, so a single "I" gets the credit.

Our proposal is that what is actually there is not one chooser. It is many settings, held at once, rising and falling: a want, a caution, a habit, a plan. What a person does is settled by their interplay (Section 7). That picture explains what the ghost was meant to explain. The steering is the person's own, because every one of those settings is a state of their own brain. It is not fixed by any one visible cause, because it comes from the interplay of many. And the person could have done otherwise in the plain sense that with a different goal in force they would have.

3 Control theory, briefly

The rest of the paper needs a handful of ideas from control theory. All of them are standard [17], and most can be felt in a shower. The idea that purposeful behavior is behavior controlled by feedback from the goal is as old as the field, which Wiener named cybernetics after the Greek word for steersman [18, 19].

3.1 The parts of a loop

Every loop has the same parts (Figure 2, Table 2). A *setting* says what is wanted. Engineers call it the set point. *Senses* report how things stand. A *comparison* takes the difference between the two, which is the *error*. A *controller* turns the error into a push. The push acts on the thing being controlled, along with every *disturbance* that also acts on it. Then the senses report again.

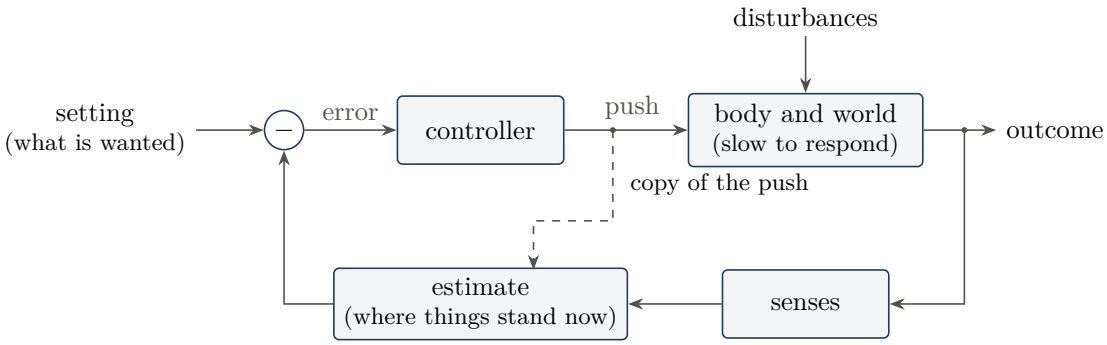

Figure 2: A loop. The comparison (the circle) subtracts the estimate from the setting. The controller turns the error into a push. The body and the world respond, late and under disturbance. The estimate combines what the senses report with a copy of the push (dashed), to say where things stand now and not where they stood when the senses last reported.

	Thermostat	Cruise control	A shower	Standing up
setting	the dial	the speed chosen	the warmth you like	upright
senses	thermometer	speedometer	your skin	inner ear, eyes, ankles
push	furnace	throttle	your hand on the valve	leg and trunk muscles
disturbance	weather, an open window	hills, wind	someone runs a tap	a shove, a moving bus

Table 2: Four loops everyone knows.

3.2 Gain: how hard to push

The simplest controller pushes in proportion to the error. Twice as far from the setting, twice the push. The factor is the *gain*. High gain makes a loop quick and stiff. Low gain makes it slow and soft. In symbols, with x for how things stand, r for the setting and K for the gain,

$$\frac{\mathrm{d}x}{\mathrm{d}t} = K\,(r - x). \tag{1}$$

This is the only equation the main text needs. Its solution creeps toward r and stays there.

3.3 Delay: the news is always old

A shower with a long pipe teaches this. You turn the valve and nothing happens, so you turn it further. Then the hot water arrives all at once. You turn it back too far. The trouble is not the valve. The trouble is that you were pushing hard on old news.

For the loop of equation (1) with a delay δ between the push and its showing up, the rule is exact. The loop is stable only if gain times delay is less than $\pi/2$, about 1.57 [20]. Below about 0.37 it settles without overshooting. In between it overshoots and rings. Above 1.57 every correction arrives too late and makes things worse (Figure 3). A slow loop has two honest choices: push gently, or stop relying on old news.

A body is a slow loop. In the human arm the spinal stretch response begins 20 to 45 ms after a push, a longer response that involves the cortex follows at about 50 to 100 ms, and voluntary reactions begin after about 100 ms [21–23]. Corrections guided by vision are slower. When the seen position of the fingertip is shifted during a reach, the hand's path changes about 160 ms later [24]. In people balancing an unstable load by hand, the hand stops tracking the disturbance above one or two cycles a second [25].

3.4 Prediction: act on where things will be

The way out is to stop waiting for the news. A *forward model* takes a copy of each outgoing command and works out where the body should be by now [26–28]. The loop then acts on that prediction, and the senses are used to correct the prediction when it turns out wrong. Engineers call the arrangement a Smith predictor [29], and it is stable at any delay (Figure 3d).

Bodies do this. When a person moves an object held between finger and thumb, the grip tightens and loosens in step with the load, with no lag [30, 31]. A reach lasting a few hundred milliseconds is stable and accurate although the feedback that guides it is 100 to 200 ms old [28]. The cerebellum has been proposed as where such predictions are made [32], and damage to it leaves movement dependent on delayed feedback [33].

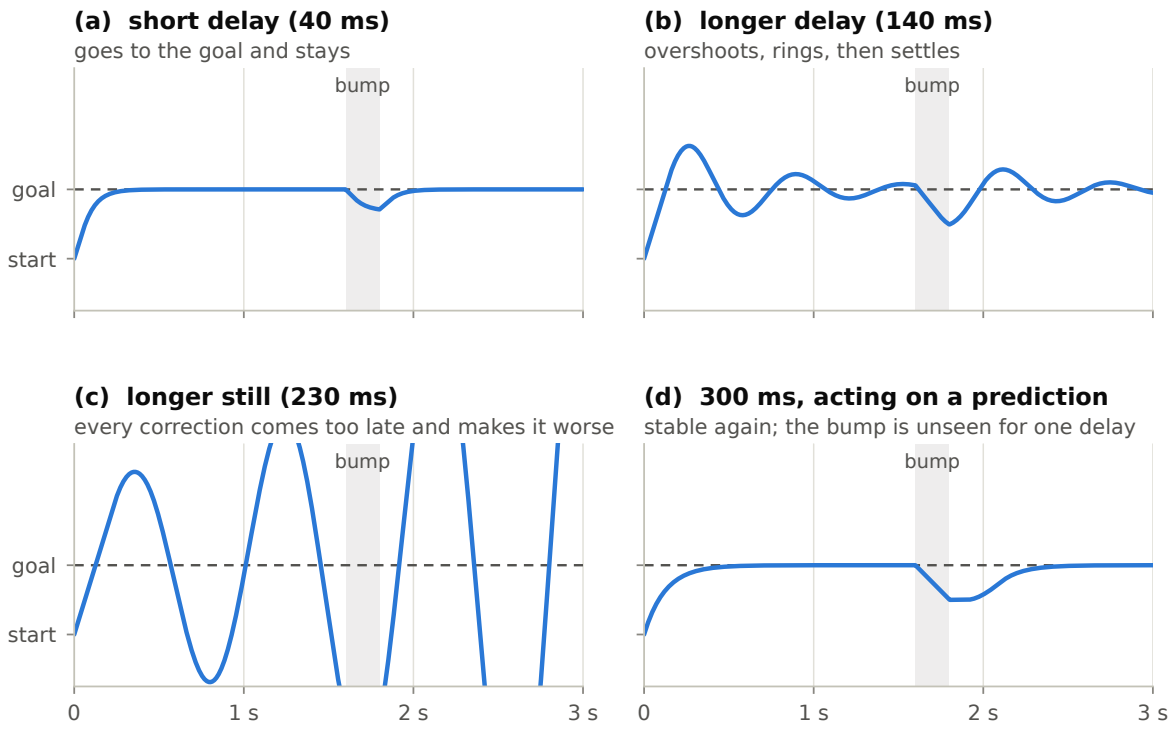

Figure 3: One loop, one gain, four delays. The setting is raised at time zero, and a bump (shaded) knocks the outcome away later. At this gain the loop stops overshooting below a delay of 46 ms and is unstable above 196 ms. In (d) the loop acts on a prediction of where things stand, and it is stable at any delay.

Perception shows the same thing. A steadily moving object is seen ahead of a flash that was physically lined up with it. Nijhawan explained this as the brain placing the moving object where it will be by the time the signal is used [34].

The consequence matters for everything that follows. *The "now" a brain acts in is an estimate.* It is not a late report. It is a prediction, checked against late reports.

3.5 Three ways to handle a disturbance

A disturbance can be met in three ways, and a body uses all three.

Predict it. If the disturbance is your own doing, cancel it as it happens. That is the grip tightening with the load. It works only when the load is the consequence of one's own action in a consistent setting [35].

Correct it. If it is a surprise, feedback is the only remedy, and it comes one delay late. After an unexpected slip the grip adjusts in 60 to 80 ms [30].

Brace for it. Stiffening a joint by tensing the muscles on both sides resists a push with no delay at all, and people learn to shape that stiffness to the task [36].

Disturbances come from inside too. Motor commands carry noise that grows with their size [37].

3.6 Loops in a stack

One loop is rarely enough. "Eat" sets "chip to mouth." That sets a path for the hand. That sets forces at the joints, and those set the pull of each muscle. Each loop's output is the *setting* of the loop below it [38]. Slow, abstract loops sit on top of fast, concrete ones (Figure 4).

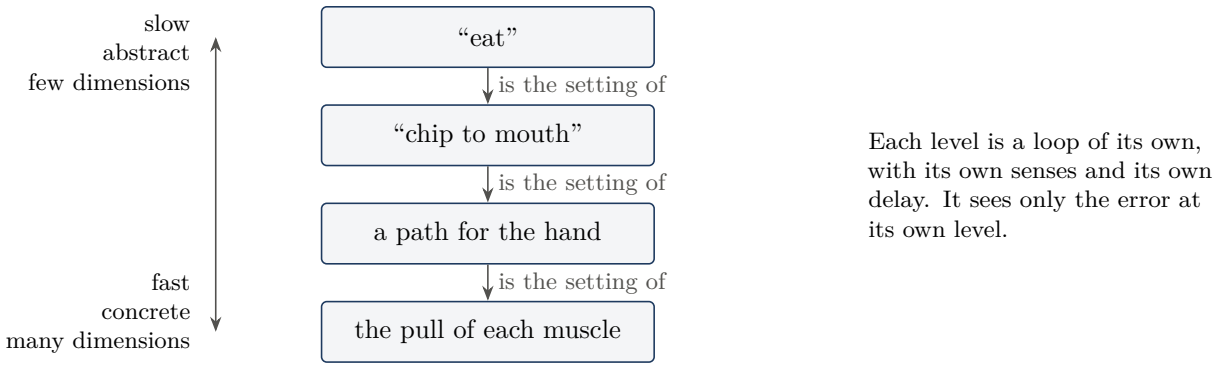

Figure 4: A stack of loops. The top says what, and each level below says more of how.

Stacked loops stay out of each other's way if their speeds are well apart. In the simplest case, an outer loop that asks for a speed and an inner loop that delivers it, the pair settles without overshoot when the inner loop's gain is at least four times the outer's (Appendix A).

3.7 Looking only now and then

A loop need not look all the time. It can take a sample, act on it for a while, and sample again. The price is that a sample held for a time h behaves like an extra delay of $h/2$ [39]. Human control of unstable loads is well described this way: watching is continuous, and action is updated at intervals, each update launching a short predicted movement [25, 40, 41]. This will matter in Section 10.

3.8 A model inside

With one dimension the sign of the error says which way to push. With many it does not. A controller facing an error in “hand position” and a body with dozens of joints has to know which joint moves the hand which way. That knowledge is a model of the body. Control theory has a theorem to this effect: a regulator that keeps working when its own parameters drift must contain, inside the loop, a model of whatever generates the disturbances and the settings it deals with [42]. A better-known result, the “good regulator theorem,” is quoted more often and proves less than its title. It shows that the simplest best regulator's actions are a mapping of the states of what it regulates, and not that every regulator carries a model it can run [43].

Word	**Plain meaning**
setting (set point)	what the loop is trying to bring about
error	the gap between the setting and how things stand
gain	how hard the loop pushes for each unit of error
delay	how long a push takes to show up in what is sensed
disturbance	anything else that acts on the thing controlled
feedback	correcting from what the senses report
forward model	a prediction of what one's own commands will do
estimate	the loop's best account of how things stand right now
stack	loops in layers, each setting the one below

Table 3: The vocabulary, for reference.

4 What a brain's setting is

A thermostat's setting is one number. A brain's is not, and the difference is the first thing to get right.

4.1 A goal fixes a few things and leaves the rest open

"Get the cup." That goal says where the hand should end up. It says nothing about the elbow, the shoulder, the tilt of the trunk or which foot bears the weight. There are countless ways to satisfy it, and the goal does not care which.

> PROPOSAL. **The setting is an abstract content.** In a brain the set point is not a single value. It is an abstracted point in a space of very many dimensions: it fixes a few of them and leaves the rest free. The more abstract the goal, the fewer it fixes and the larger the set of states that satisfy it. And it is held in the same format as understanding.

Figure 5 shows the smallest case. An arm with three joints reaches for one target from four starting poses. The hand ends in the same place every time. The arm does not. The goal is not a point the arm must reach. It is a whole curve of arm positions, every one of which puts the hand on the target, and the loop pulls the arm to the nearest part of it.

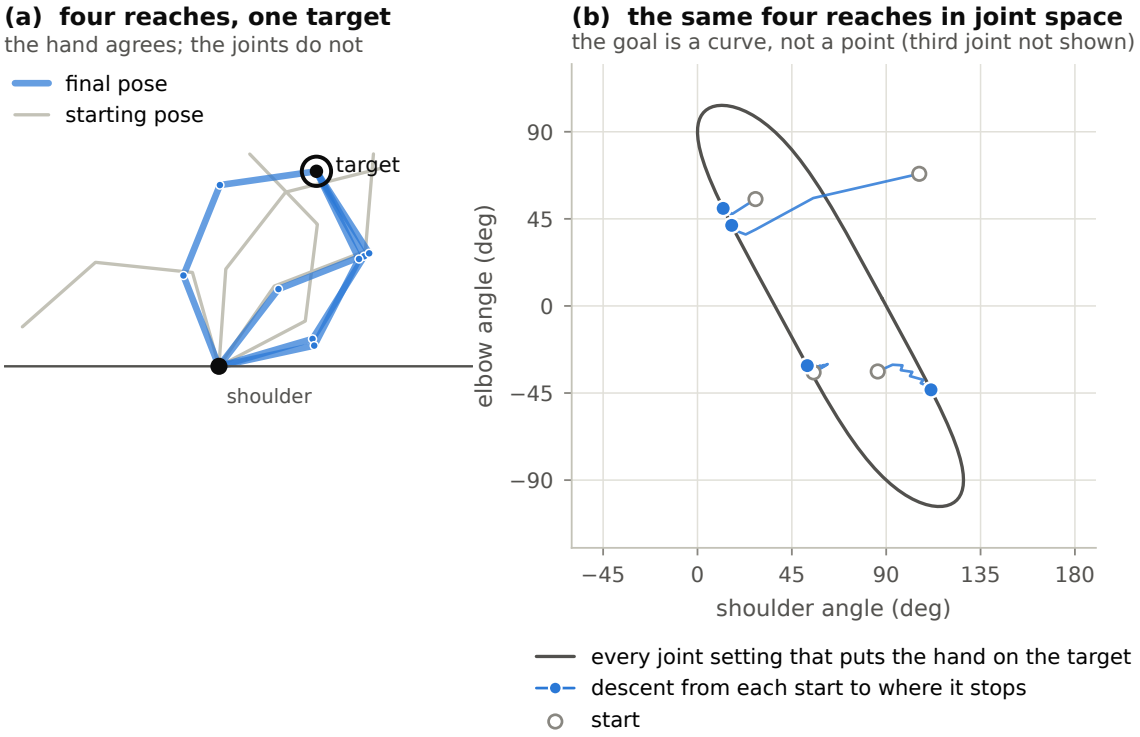

Figure 5: A goal with fewer dimensions than the body. (a) Four reaches from four starting poses (gray) to one target. The hand agrees every time and the joints do not. (b) The same reaches drawn by joint angle. The goal is a closed curve: every point on it puts the hand on the target. Each reach ends on the curve at a different place.

Real movement looks like this. Bernstein saw that the body has far more ways to move than any task needs, and called that the central problem of coordination [44]. When people repeat a

movement, their joints vary more in the ways that leave the task unchanged than in the ways that would spoil it [45]. A controller that saves effort would do exactly that: correct a deviation only when it interferes with the task [46]. The variability is not sloppiness. It is what a goal that fixes only a few things looks like from outside.

4.2 A goal has to be in the same terms as the estimate

A comparison can only subtract like from like. In the fruit fly of Section 5, the goal and the heading are both written as directions in the world, and that is what lets one set of cells compare them. A person's estimate of how things stand is in sights, sounds and feels. So we propose that a person's goal is held in those terms too, as a *would-be percept*: what it would look and feel like to have it. The error is the difference between that and what is actually perceived, and action is whatever makes the difference shrink.

Several lines of work say something like this in their own words. Perceptual control theory holds that what an organism controls is its perceptual input, brought to reference values set inside it [38]. Active inference treats the signals that descend to the muscles as predictions of what the body will feel, which reflexes then make true [47, 48]. The theory of event coding holds that actions are represented by their perceivable effects, in a code shared with perception [49]. Ring Bank Theory describes top-down content as "a template summoned to meet the world" [16, §1.9], which is a description of a setting.

4.3 A goal takes a moment to set

> Observation. **A goal arrives as an image that is understood.** A mental image may take a second or more to set, to understand and to focus on. The understanding then has the set point in it, as part of it. After that the means have to be determined (driving to a particular restaurant, say), and the plan comes to mind.

The wait has a function. A setting is useful only if it holds still.

> Proposal. **A goal must settle before it is pursued.** A content can serve as a setting only once it has settled, and it must then hold still for much longer than the loops beneath it take to run. A loop handed a setting that is still forming steers toward wherever the setting was a moment ago. The second or more of the observation above is the settling.

The book gives the time for content to become understood as tens of milliseconds for familiar things seen and as several hundred milliseconds, or seconds, for imagined and abstract content [16, §1.3]. A goal is of the second kind.

Once the goal has settled, the stack of Section 3 fills in the how. For an abstract goal the top of the stack is a plan, and the classical account of planning is means–ends analysis: find the difference between where things stand and the goal, choose an action that reduces it, and if that action cannot be taken yet, make whatever it lacks into a goal of its own [50]. "Have dinner" becomes "be at the restaurant," which becomes "get the keys."

5 How a brain holds a set point

A thermostat holds its setting in a dial. Where does a brain hold "get water," or "walk that way," or "be at the restaurant by seven"?

Two answers suggest themselves. One is a *node*: cells wired for that goal. The other is *inertia*: a pattern of activity that a network was set doing and keeps doing. This section goes through the cases in which the cells are actually known. Brains turn out to do it in at least three ways (Table 4), and one step is still not understood.

	No dial	**Need cells**	**A held pattern**
example	body temperature	thirst, hunger	a direction to walk in, a planned move, a place to get to
where the set point is	nowhere: several reflexes, each with its own trigger, and the body settles where they balance	no stored target: the cells fire when the body is short	a pattern of activity that stands for the target
what keeps it going	the tuning of the sensors	the shortage itself	the circuit's own activity
what ends it	each reflex stops on its own	word from the mouth and gut that relief is coming	reaching it, a signal to go, or forgetting
node or inertia	neither	node	inertia, on wiring that gives it meaning

Table 4: Three ways a brain holds a set point. The sorting is ours. The findings behind each column are cited in the text.

5.1 No dial: body temperature

The body's own thermostat has no dial. On Romanovsky's account, body temperature is not regulated by one controller with one set point. It is regulated by several independent loops, such as shivering and sweating [51], each with its own sensors and its own trigger temperatures. Nothing adds the readings into one body temperature, and nothing compares that with a stored target. The level the body keeps is where the loops balance. He proposes calling it a balance point, and dropping the term set point for it [52].

This still meets the definition of steering in Section 1. Shift a trigger and the level the body keeps shifts with it. But there is nothing here that could be called a goal the animal has. Steering does not require one.

5.2 Need cells: thirst and hunger

Thirst has cells. In the mouse they sit in a small structure of the forebrain, and they fire when the body is short of water. They do not hold a target that the body's state is checked against. Their firing is the shortage. More exactly it is the shortage as predicted: single cells add up what the blood reports with what the mouth and the gut report [53].

How a need steers. It does not tell the muscles what to do. It changes what the rest of the brain does with what it senses. In recordings from about 24,000 neurons across 34 regions, a cue that predicted water set off activity that spread through the brain of a thirsty mouse. The spread was gated by a brain-wide pattern of activity that carried the state of thirst, and switching the thirst cells back on after the mouse had drunk its fill brought the thirsty pattern back [54].

What ends it. The thirst cells are quieted in three stages, fastest first. Drinking itself does it: a group of inhibitory cells comes on the moment the mouse starts to drink and shuts the thirst cells

down, and without that group the mouse drinks too much [55]. Then the gut reports how watery or salty the intake is [53]. Last, the blood changes. The first two arrive long before the blood is corrected, and they have to, because most drinking is over too quickly for the blood to control it [56]. This is the delay and prediction of Section 3, found inside a need.

Hunger has the same design. Switching on one type of cell in the hypothalamus makes a mouse eat within minutes, with no training, and more of the cells means more eating [57]. Those cells come on with a shortage of energy and are rapidly turned down when food is detected, before any is eaten [58]. Mice avoid having them switched on [59]. Calories arriving in the stomach turn them down quickly and lastingly, in proportion to the calories, and the hormone leptin acts only over hours [60].

So a need is a node. Specific cells fire until relief is predicted. No target is written anywhere.

5.3 A held goal, cell by cell: the fly

A walking fruit fly will hold a straight course along one bearing for hundreds of body lengths [61]. The cells that do this are known, and they make a complete loop (Figure 6).

The estimate. The fly's heading is a bump of activity in a ring of cells. The ring keeps the bump going by itself. When experimenters wrote in an artificial bump, the ring maintained it [62].

The set point. A second set of cells carries the direction the fly is trying to go. Switching them on at a chosen place makes the fly walk in the direction the experimenter picked [63].

The comparison. A third set receives both signals, and each of its cells responds to both. The group on one side is active when the fly points to the left of its goal, and it drives a turn to the right. The group on the other side does the reverse. A further set turns the steering up when the fly is far off course [64].

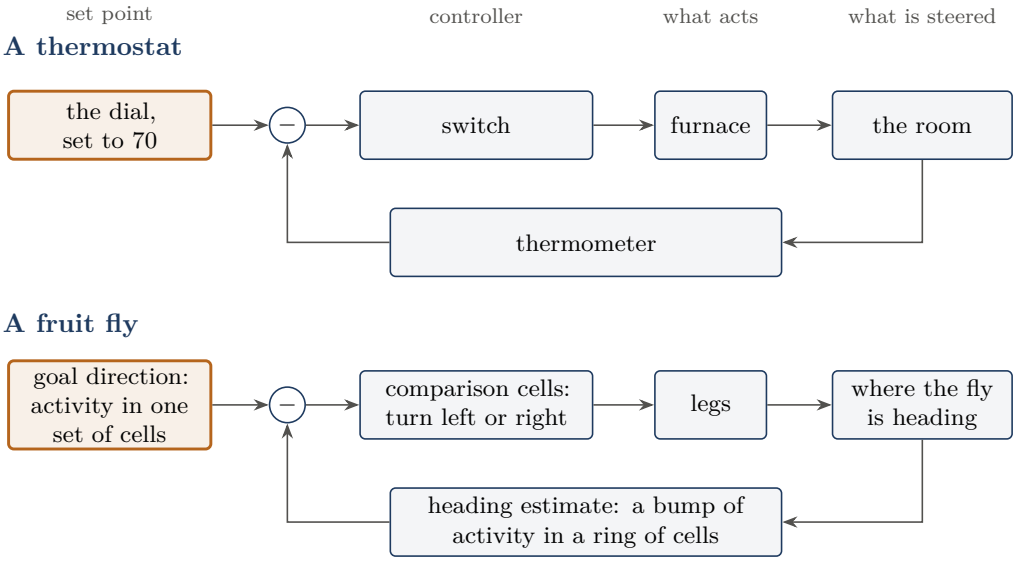

Figure 6: The same loop twice. In the fly every box is an identified class of cells: FC2 for the goal, EPG for the heading, PFL3 for the comparison and the turn [63, 64].

Three things in this circuit matter for the rest of the paper.

The fly steers by its estimate. When experimenters rotated the heading bump artificially, the fly slowed and turned so as to bring the bump back to where it had been [61]. It was correcting its

own estimate, which is the wiggle test passed by a real brain.

Goal and estimate are written in the same terms. Both are directions in the world, and that is what lets one set of cells compare them and turn the difference into a turn of the body [63].

The goal can be any direction. Blocking part of the comparison cells made flies worse at holding arbitrary goal directions, by the amount the wiring predicts [63]. One fixed circuit serves every goal of this kind. The wiring gives each position in it a meaning, and the activity picks one. Here the node answer and the inertia answer are both true at once.

5.4 What keeps a goal alive

A goal held over a delay has to be kept going by something. Three cases are known, and they differ.

Two areas keeping each other active. A mouse that must wait before licking left or right holds the plan in a region of frontal cortex. That region and a part of the thalamus excite each other. Silencing the thalamus makes the cortical activity collapse almost completely and at once, and silencing the cortex cuts the thalamic activity [65]. This is a loop of neurons in the literal sense. Its job is to hold the plan. The loop that does the steering is the other one, out through the body and back through the senses.

The network, and no single cell. In the goldfish, cells that hold the eyes still after each movement keep up their firing because of sustained input from the rest of the network. A single cell pushed with current only changes briefly [66].

Bursts, with synapses in between. Early recordings found prefrontal cells in monkeys that change their activity through the delay between a cue and the response to it [67], and working memory came to be thought of as steady firing. Miller's group found something else. The firing that carries a remembered item comes in brief bursts, tied to bursts of fast rhythm, when the item is put in and when it is read out. A slower rhythm fills the time between [68]. Between bursts the item is thought to be carried by short-lived changes at the synapses just used [69–71]. The slow rhythm is tied to top-down signals and to inhibition, and it regulates the fast one, which is tied to what comes in from the senses. On this account that interplay is how what is held gets controlled [71].

What is held there can be abstract. In monkeys trained to apply a "same" rule or a "different" rule to pictures they had never seen, the commonest activity among prefrontal neurons coded the rule [72]. On Miller and Cohen's account that is the job of prefrontal cortex: to keep patterns going that stand for goals and the means to them, and so to tilt activity elsewhere toward what the task needs [73].

On every one of these accounts a goal lasts only while something keeps it up, which is why a goal of this kind can simply lapse. Items that are not rehearsed leave short-term memory within seconds to tens of seconds [74, 75]. A habit is different. Routines made stereotyped by learning run with little deliberate control [76], and they are learned by other machinery than the memory of events [77]. A habit's goal does not have to be kept alive. The situation calls it up.

5.5 Does the held pattern do the steering?

In each case where it has been tested, yes.

In the fly, switching on the goal cells at a chosen place sends the fly that way. In the mouse, knocking the held activity off course usually changes nothing, because it settles back. Now and then it settles into the state for the other movement instead, and then the mouse makes the other

movement [78]. In the rat, cells in orbitofrontal cortex point to the place the animal is about to go to. They do so from before it sets out, with the goal out of sight, and on trials where the rat goes to the wrong place they point to that wrong place from the start. Disrupting them briefly at the start of a journey sends the rat astray [79].

Holding a plan and launching it are separate. The mouse's held plan waits for a signal to go, which travels from the midbrain through the thalamus and flips the cortex from holding the plan to commanding the movement [80].

5.6 Any goal? What is known, and what is not

A person can make nearly anything a goal. Some of what makes that possible is known.

The machinery that holds a goal is not built for particular goals. A new goal is a new pattern over cells that already exist. Single prefrontal cells respond to mixtures of the features of a task, which greatly enlarges what can be read from the population [81]. Miller's group proposes that the control of what is held, such as the order of items, is carried by where in the network the activity is, apart from the connections that store the items themselves, and that this is what lets the same control apply to new content [82]. Routes to a goal can be composed on the spot: before a rat sets out, its hippocampus plays through paths from where it is to the place it means to reach, including for pairs of start and goal it has never used [83].

One step has not been shown. The fly has cells built to compare one kind of goal with one kind of estimate. A brain like ours cannot have cells built for every goal a person can think up. How it checks a made-up goal against how things stand is not established in anything we have read.

> PROPOSAL. **A made-up goal is checked by perception run backwards.** A goal is held at the top of the network that perceives, and is fed downward as what the world would look and feel like if the goal were met. The mismatch between that and what the senses report is the error, and the loops below act to cancel it. The comparison uses machinery that perception already has, so it works for anything that can be imagined.

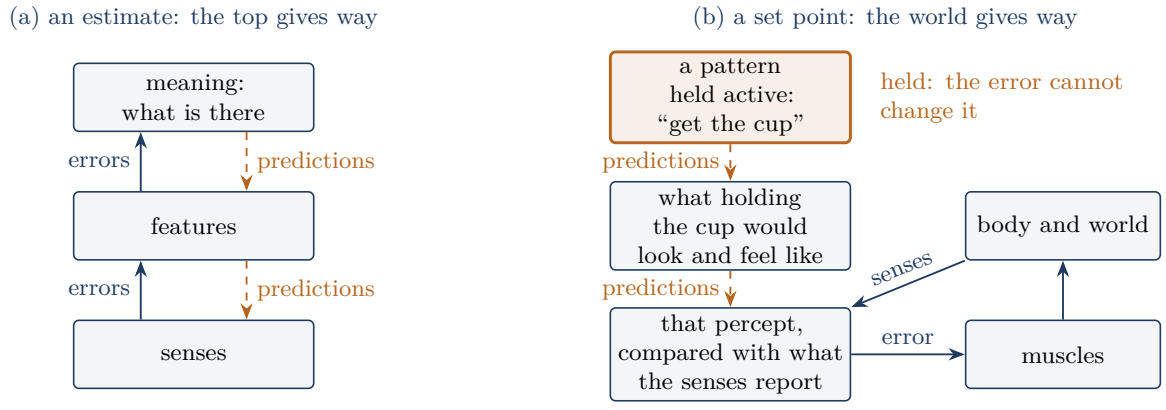

Figure 7: The proposal. (a) Perceiving. Predictions come down, errors go up, and the errors revise the top, which comes to say what is there. (b) Wanting. A pattern at the top is held. Its predictions are what the goal would look and feel like. The top does not give way, so the error is canceled by action.

The machinery is real for perception. In a model of visual cortex in which feedback connections carry predictions of the activity of the level below, and feedforward connections carry the errors, model neurons shown natural images developed receptive fields like those of the simple cells of visual

cortex [84]. Active inference keeps that wiring and changes which side gives way: the descending signal is a prediction of what the body will feel, and reflexes make it true [48]. On this picture an estimate is content that yields to the senses, and a set point is content that is held so that the world is made to yield.

It is a proposal. Nothing we have read shows it for goals. And it leaves a second step open: run backwards, a network produces what the goal would look like. It does not say which action gets there. That is the work of the stack of Section 3 and, for abstract goals, of planning.

5.7 So: node, or inertia?

Needs are nodes. A chosen goal is inertia, and what the inertia means is fixed by the wiring it runs on. The reflexes that keep the body's temperature need neither.

To tell whether some activity is a set point at all, and not a plain command, hold it fixed and vary the circumstances. If it is a command, the movement repeats and the outcome varies. If it is a set point, the outcome repeats and the movements vary. The fly's goal cells come closest: with their activity set by the experimenter, the fly ends up heading the chosen way. For most of the rest, only half the test has been done.

In Ring Bank Theory's terms, the variable for intent is the aboutness vector, a direction in concept space that says what the system is about. It "supplies intent," and in its top-down form it is "self/goal-driven" [85, §2.9, §4.7]. A pattern that holds a set point holds a place in that space.

6 Kinds of goals

A thermostat has one kind of goal: keep one number at one value. A person has many kinds, in force at the same time. Sorting them is worth the trouble, because the kinds are served differently, kept differently and broken differently.

6.1 Shapes in time

A goal can be sorted by the pattern over time that satisfies it [86, 87]. Figure 8 shows six shapes, each met and each not met.

The six fall into two families, and the difference is practical. A *never* goal says that something bad never happens. An *eventually* goal says that something good sooner or later does [88]. Computer scientists call them safety and liveness, and every requirement on behavior over time is a combination of the two [89].

A never goal can be broken, at a moment you can point to: the cup spilled *then*. An eventually goal is never broken. At any moment it is only unmet so far. So a never goal needs watching, and an eventually goal needs progress. Planning has the same split under the names maintenance and achievement [90].

6.2 Five more things to know about a goal

Shape is one entry on a goal's card. There are five others.

> Proposal. **A goal's card.** Each goal in force at a given moment is described by six entries.
>
> **About** the few things it fixes (Section 4).
> **Shape** its pattern in time (Figure 8).

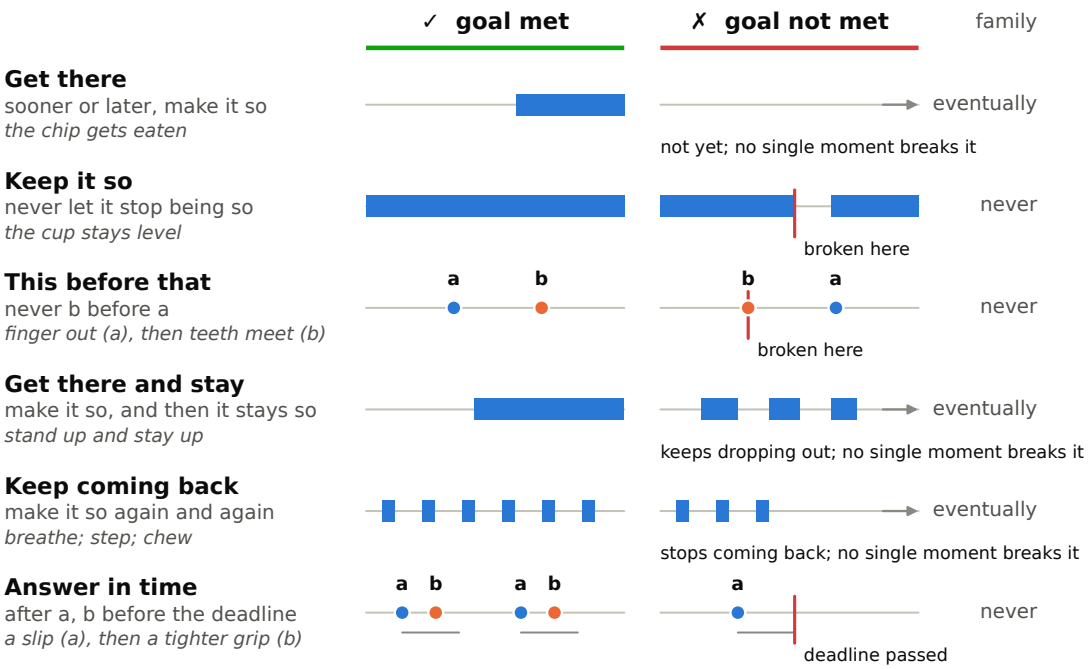

Figure 8: Six shapes a goal can have in time, with one timeline that meets each and one that does not.

Standing	whether it can start behavior or can only restrain it.
Binding	whether it is a must or a preference to be traded, and what it outranks.
Clock	how fast the loop that serves it runs.
Kept by	what holds it, which fixes how long it lasts: a reflex, a habit, or working memory.

Standing needs a word of its own, because it is what people usually mean by a "weak" goal.

DEFINITION. **Standing.** A goal has *initiating standing* if it can start behavior. A *restraint* is a goal that cannot. It acts only on behavior that another goal has started, by delaying it, blocking it or shaping it.

"Don't spill" never got anyone out of a chair. It only shapes how the walking is done. Engineering has the same division. A supervisor in supervisory control can forbid events and cannot force any event to occur [91]. A control barrier function leaves a controller's command alone unless that command would carry the system out of a safe set, and then alters it as little as possible [92].

6.3 A worked case: the flash while eating chips

OBSERVATION. **A flash image while eating chips.** While eating chips there is sometimes a flash image of biting one's own finger. The image is made in about a millisecond, persists visually for about 100 ms, and persists as understanding in working memory for a few seconds. A threat or an affordance can be present in such an image as common sense. The caution does

not come with the flash. There is a little delay between the flash and the "decision" to be careful about when to bite down: make sure the finger is out. The result is a weak set point, "don't bite the finger," alongside the set point "eat chips." Three more things are part of the observation. (i) The restraint is about *when* the jaw closes: the finger must be out before the bite. (ii) "Weak" means that the goal is subservient, a restraint. It does not mean that it weighs little or that the margin it keeps is small. (iii) It fades because a little later one may forget to be careful.

The durations are reports in perceptual time. Every entry of the card shows up in this one small event, so we take them in turn.

Where the image comes from. The forward model of Section 3 predicts where the body is now. Nothing stops it from running a little further ahead, along the movement in progress.

PROPOSAL. **The flash is a prediction, and the caution follows when it is understood.** The flash is the forward model run ahead of the present, surfacing a state the plan would pass through: a finger between closing teeth. The threat is in the image, as common sense. The restraint is not in force at the instant of the image. It follows after the little delay of the observation, when the image has been understood as a threat. A predicted state understood as a threat is a restraint. Understood as something good, it is a thing to bring about.

Ring Bank Theory lists among the jobs of imagined content "warning imagery (e.g., potential bitten-finger flash when bringing fingers into mouth while eating chips to induce caution)" [85, §2.3]. It also has the delay. An event is registered in under a millisecond and may not be understood for half a second [16, §1.3], and content is first formed and then read for threat and affordance, "shaping behavior" [85, §7]. Engineers build the same operation on purpose and call it model predictive control: predict the system's states over a horizon, check them against the constraints, and change the plan before a predicted violation happens [93].

A single image can carry a goal only when the goal has one defining moment: a bad moment to avoid, like the bitten finger, or a good one to bring about, like the table at the restaurant. Goals with no defining moment (keep breathing, keep the cup level) cannot be carried by one image. They need a rhythm, a rule or a standing habit.

What the caution changes. It changes *when*. Figure 9a shows a bite as two movements: the finger coming out and the jaw closing. In the model the finger is clear 170 ms after it starts to move, and the teeth meet 127 ms after the jaw starts. So the jaw has to start at least 43 ms after the finger. A habit that starts it 100 ms after leaves 57 ms to spare. The caution moves that one event later, to 150 ms, and leaves 107 ms. Neither movement changes shape. In the theory of systems that mix events with continuous motion, a condition that must hold before an event may occur is called a guard [94]. "Finger out" is a guard on "teeth meet."

The margin is there because timing is noisy. If the jaw's start varies by 25 ms from bite to bite, the habit's 57 ms margin fails about once in 90 mouthfuls, and the careful margin fails about 9 times in a million.

Why it is "weak". The caution starts nothing. Nobody eats a chip in order to not bite a finger. It rides on "eat chips" and delays one event of it. That is a restraint in the sense of the definition, and it is weak in standing, not in force. While it holds, it is a must.

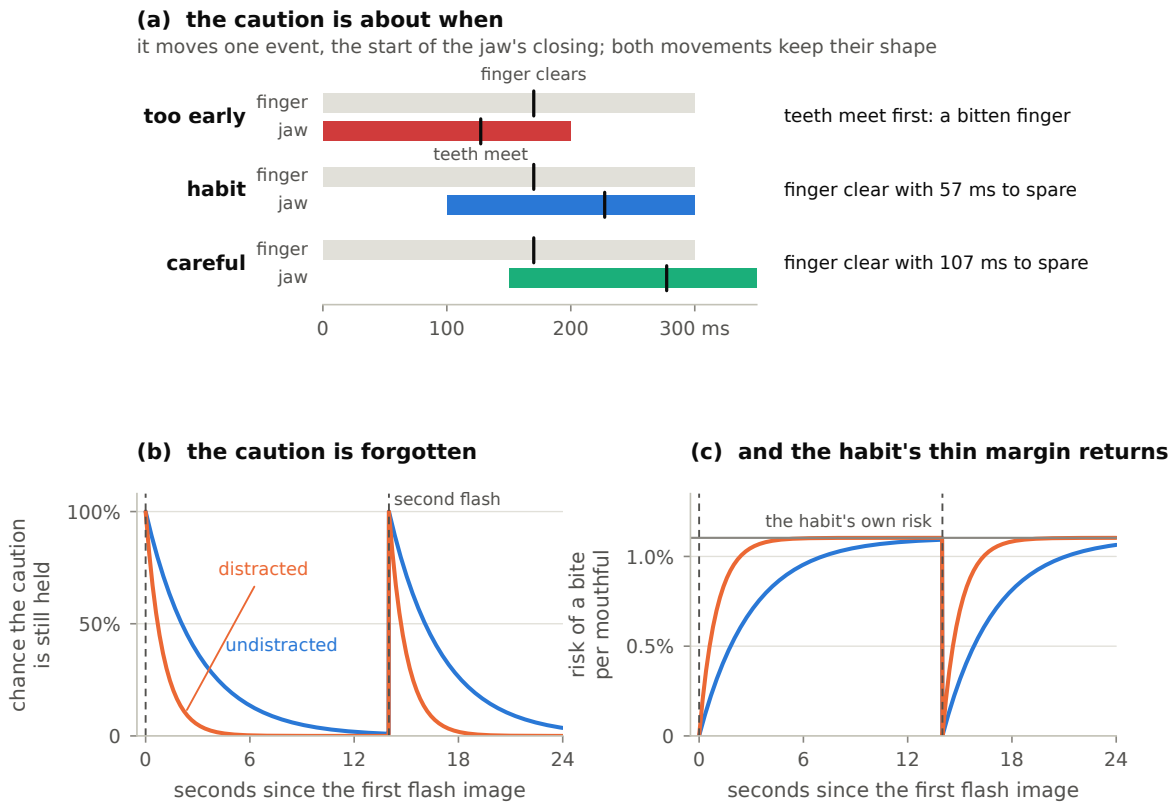

Figure 9: The caution, in a toy model. (a) Three timings of one bite. The gray bar is the finger's withdrawal and the colored bar is the jaw's closing. Ticks mark the finger clearing and the teeth meeting. (b) The chance that the caution posted by a flash is still held, for a memory that lasts about 3 s, or 1 s under distraction. A second flash posts it again. (c) The risk of a bite per mouthful, if timing varies by 25 ms from bite to bite. The numbers are chosen to illustrate and are not measurements.

Why it fades. The caution does not ease off because the danger has passed. It is forgotten.

> PROPOSAL. **The caution is held, not installed.** The restraint of the chips observation is not written into the eating habit. It is an override on the habit, held in working memory. Working memory is volatile. When the override lapses, control returns to the habit, with whatever margin the habit had before.

The standard account of slips of action has this shape. Routine action runs from habit without deliberate control. A supervisory system that needs attention can bias which routine runs, and when attention goes elsewhere the strongest routine takes over [95]. Lost intentions are a recognized class of slip [96, 97]. People can state a requirement of a task and still fail to act on it, most often when several requirements are new at once [98].

Figure 9b and c put this into the model. If the caution lasts about 3 s in memory, the chance it is still held is 0.51 after 2 s, 0.19 after 5 s and 0.04 after 10 s, and the habit's risk comes back at the same pace.

> HYPOTHESIS. **How long, and what would make it last.** A caution that arrives as a flash image lasts as long as its understanding stays in working memory, the "few seconds" of the

observation, unless the flash comes again and posts it again. A lasting change to the habit itself takes an experienced error, a real bite, and does not come from the image alone.

The second half rests on an analogy with learning to move. What updates a movement from one try to the next is the error experienced on the last one [99, 100].

6.4 Many goals, one body

Now a busier case: carrying a full cup across a room while talking. Table 5 fills in the cards.

Goal	Shape	Standing	Binding	Clock	Kept by
Eating chips					
chip eaten	get there	starts	preference	∼1 s	habit
finger unbitten	this before that	restrains	must	one bite	working memory
Carrying a full cup across a room while talking					
reach the far side	get there	starts	preference	seconds	working memory
stay upright	keep it so	starts	must, first	50–100 ms	reflex and habit
step	keep coming back	starts	preference	∼0.5 s	habit
cup level	keep it so	starts	must	∼200 ms	working memory
no gesture with that hand	keep it so	restrains	must	∼200 ms	working memory
finish the sentence	get there	starts	preference	∼200 ms	working memory

Table 5: Goal cards for two everyday cases. Clock values are orders of magnitude from the loop delays of Section 3 and from ordinary experience. They are not measurements.

Six goals share two hands and one pair of legs, and they do it in four ways.

PROPOSAL. **Four ways goals share one body.**

Different clocks. A fast loop inside a slow one does not fight it. Balance corrects in a tenth of a second and the walk across the room takes many seconds.

Different dimensions. One goal leaves freedom spare and another spends it. The lower-ranked goal acts only in ways that leave the higher one untouched. Robotics calls this task-priority control [101]. It is the arm of Figure 5 put to work.

A trade. Goals that want the same thing differently, and are only preferences, are folded into one cost, and the body takes the compromise [46]. Both goals bend the whole movement.

A veto. A restraint does not enter the trade. It inspects what the other goals are about to do and delays or blocks the part that would break it [91, 92]. Far from a violation it changes nothing.

A trade and a veto can be told apart from outside. A trade bends the whole movement toward a compromise. A veto leaves the movement alone and acts sharply at the one place where the violation would have been. In the chips model the caution moved one event and both movements kept their shapes, which is a veto.

The cup uses all four. Balance outranks everything and runs fastest. The arm with the cup gives up its swing, which is spare freedom spent on keeping the cup level. “No gesture with that

hand" vetoes what talking would otherwise do. And the cup and the sentence both live in working memory, where doing two attention-demanding things at once has a measurable cost [102].

That a brain holds several possible actions at once has support from recordings. On the affordance competition hypothesis the brain specifies several potential actions in parallel, and they compete until one is released [103].

> Hypothesis. **Only a few new cautions at once.** Restraints held in working memory compete for its capacity, which is about four items [104]. Restraints carried by reflexes or by habits do not. So a person can keep only a few new cautions in force at once, and there is no such limit on practiced ones.

7 Deciding without a decider

Section 6 ended with six goals in force at once and nobody in charge of them.

> Proposal. **The range of freedom.** At any moment some regions or ensembles of a brain speak up, and each stands for the meaning of a set point: a want, a thought, a habit, a caution. They rise and fall. Their interplay settles what is done, and taken together it is the person's decisions and the person's personality. No further party listens to them and picks. What is done is laid down as history, by the hippocampus and by slower learning. A person's freedom has the range of the set points their brain can voice.

7.1 The interplay has rules

The interplay is governed by rules. Goals share the body by the four rules of Section 6: by speed, by dimension, by trade and by veto. Possible actions are specified several at once and compete until one is released [103]. Habits contend for control, and a supervisory system that needs attention tilts the contest [95]. On the account of prefrontal cortex in Section 5, that tilt is what a goal held in mind supplies [73]. Each of these is a mechanism.

7.2 One voice at a time, many goals in force

Experience does not feel like six goals at once. It feels like one thing after another. Ring Bank Theory has both. Its Mic foregrounds exactly one high-level voice at a time [16, §1.3]. Experience is "monophonic at the top even when many things are present: one voice sounds, however many keys are down" [16, §1.9]. The voice that sounds and the goals in force are different things (Figure 10). A goal is posted when its voice has sounded and been understood. After that its store keeps it in force, a habit for years and working memory for seconds, while other voices sound. "Stay upright" is in force all day and almost never has the floor.

7.3 A decision, and a personality

A *decision* is what the interplay settles on one occasion: the moment one goal's claim on the body wins out and the others give way or wait. A *personality* is what stays the same in the interplay across occasions: which goals a person's brain voices readily, which ones outrank which, how long a caution lasts before it is forgotten. Both words name the same machinery, at two time scales.

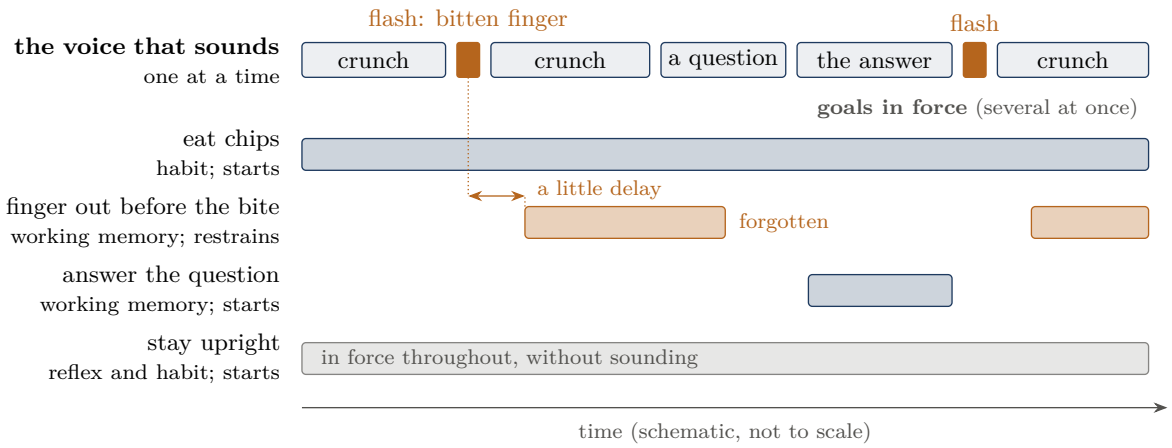

Figure 10: One voice, several goals. A schematic, with nothing in it measured. Top lane: the voice that sounds, one at a time. Below: the goals in force, several at once. A goal is posted when its voice has sounded and been understood, and its store keeps it while other voices sound. The caution follows its flash after a little delay, and is forgotten a few seconds later.

7.4 The range, and how it grows

A person can steer toward whatever their brain can voice as a goal, and toward nothing else. That is their range. It differs from person to person, and it is not fixed.

> PROPOSAL. **Set points can be invented and tried.** The range is not fixed. One can invent, imagine or ruminate one's way to ideas, goals and set points. And one can play along with a set point to try it out, a team's or a group's for instance, and see how it goes.

In the loop's terms, imagining is the forward model run on purpose. It is the operation that produced the flash over the chips unasked, and Ring Bank Theory lists hypothetical situations and action simulations among the jobs of imagined content [85, §2.3]. Trying a goal on is holding it provisionally, while a slower loop watches how things go and keeps it or drops it. A person can adopt a goal they did not start with, and the adopting is itself steering.

7.5 History

What is done is laid down, and what is laid down comes back. Damage to the hippocampus on both sides is followed by a loss of recent memory [105], and habits are laid down by other machinery (Section 5). The goals and habits that speak up tomorrow are in part the record of what was done today. That is how a personality gets made, and it is why trying a goal on can change who tries the next one.

8 So is this free will?

The short answer is that it is real steering, it is yours, and it is no escape from physics. Whether to call that free will depends on what the phrase was supposed to promise. This section sorts out what the loop gives a person, what it does not, and where that leaves the old argument.

8.1 What you get

You get steering, in the exact sense of the definition in Section 1. What you do follows goals that your brain holds, and it follows them through disturbance: the detour, the interruption, the bad night. The goals are states of your own brain. There are many of them and they are yours: your wants, your cautions, your habits, your plans. You have a range, and you can widen it by inventing goals and trying them on. Holding a goal against the pull of the moment is what Miller and colleagues describe as breaking free from reflexive reactions to gain control over one's own thoughts [71]. We call the whole of this *self-determination*.

8.2 What you do not get

You do not get to be the first cause of yourself. Something set each of your goals. A higher goal did, and above the highest there are learning, inheritance and the history of the species. Even the knack of inventing a goal was not invented by you. No loop draws its topmost setting from nothing.

Strawson built a general argument on this. To be ultimately responsible for what you do, you would have to be responsible for the way you are, and nothing can be the cause of itself [106]. The loop picture grants the premise. Control gives self-determination. It does not give ultimate origination.

8.3 Three readings

Whether self-determination deserves the name "free will" is the old dispute, and each side has a serious statement. We claim self-determination and no more.

"That is free will". On the compatibilist reading the freedom worth wanting was always a kind of control, and determinism was never its opposite [107]. Fischer and Ravizza's version is the closest to our terms. Responsibility requires guidance control: the act comes from a mechanism that is the agent's own and that responds suitably to reasons, and this is compatible with determinism [108]. The definition of steering is that idea in control terms, and this is the reading our argument supports.

"That is not enough". On the hard determinist reading, if determinism is true then our acts are consequences of the laws of nature and of events long before we were born. Neither of those is up to us, so neither are their consequences [109]. Pereboom adds that a person whose goals were fixed by causes beyond their control does not differ in the way that matters from a person whose goals were installed by a manipulator [110]. The regress of the last subsection is exactly what this reading points at.

"Free will needs a gap". On the libertarian reading free will is real and requires that determinism be false. Kane defines it as the power to be the ultimate creator and sustainer of one's own ends, and puts the needed openness in choices made when a person is torn between purposes that are both their own [12].

What the loop adds. It cannot settle this. It can say three things that bear on every side. First, an undetermined event downstream of a person's goals is a disturbance, and it makes what they do depend *less* on what they want. So a libertarian account has to put its openness among the person's own goals, as Kane's does, and nowhere after them. Second, the ghost of Section 2 is one libertarian

mechanism and not libertarianism. Refusing it leaves all three readings standing. Third, the range of freedom is common ground. On any reading, what a person can steer toward is bounded by the goals their brain can voice.

8.4 Objections

If it is all determined, why bother deliberating? Because the deliberating is part of what determines it. Apply the wiggle test. Take away the weighing of one goal against another and the outcome is different. So the outcome depends on the weighing. A future that is fixed *by way of* your deliberation is not a future that would arrive without it.

Could I have done otherwise? In one sense yes and in another no, and the three readings disagree about which sense matters. Yes: with a different goal in force you would have done something else, and which goals are in force is settled inside your own brain. No: with every particle the same, the same thing happens, unless chance steps in, and chance is not control.

If my brain is determined, could someone tell me now what I will choose? Not in a way you would be bound by. A prediction handed to you becomes one more thing your brain takes into account, and it may undo itself. MacKay showed that no prediction of a person's choice exists that the person would be correct to accept before choosing [111].

Would a little randomness make me freer? No. In a loop randomness is noise. A coin toss in the middle of a reach is a shove on the arm.

Does a thermostat have free will, then? It steers, and the definition is meant to include it, since steering needs nothing beyond ordinary physics. What it lacks is range. A brain holds abstract goals, holds many at once, sets them itself through higher goals, invents new ones, and steers by a model that includes its own body.

Does this explain consciousness? No. It says what a conscious model would be *for*: it is the estimate that the slow, goal-setting loop runs on (Section 10). It does not say why holding such an estimate is felt. That is the hard problem [112], and nothing here touches it. Attention schema theory also derives awareness from a control model, the brain's model of its own attention [113], and it too is an account of function. Ring Bank Theory says of itself that it "does not claim to dissolve the felt character of experience into geometry" [16, §1.9].

9 Changing how the loop runs

So far the loop has been taken as it stands, and the question has been what its goals do. Behavior can also be changed without touching any one goal, by changing how the loop runs. A thermostat has such settings too: how far the temperature may drift before the furnace comes on, and how hard the furnace runs. Change them and the house behaves differently with the dial untouched. Age, sleep and chemistry change a brain in this way.

> Proposal. **The loop has settings.** Besides its goals a loop has settings, which fix how it runs whatever the goals are. Four matter here: *plasticity*, how fast experience rewrites goals, models and habits; *gain*, how hard an error is acted on; *breadth*, how far behavior strays from the best-practiced option; and *hold*, how long a goal or a caution stays in force once it is posted.

Setting	More of it buys	More of it costs	What moves it
plasticity	fast learning, and a wide range of what one could come to want	precision, and whatever the world presents gets written in	high in childhood; raised for a time after some drugs, in animals
gain	speed and decisiveness	overshoot, when the news is old	dopamine raises it
breadth	discovery	reliability	high in childhood; psychedelics, on one model
hold	follow-through, and cautions that last	flexibility	sleep loss lowers it; so does blocking NMDA receptors

Table 6: Four settings of the loop. The first three columns are the proposal. The last is a hypothesis, built on the studies cited in this section.

The brain has machinery of this kind. On one computational theory each of four ascending neuromodulators sets one parameter of learning for the whole brain: how wrong a predicted reward was (dopamine), how far ahead rewards are counted (serotonin), how random the choice of action is (noradrenaline), and how fast memory is overwritten (acetylcholine) [114]. And at least one setting is tuned to circumstances. People give recent outcomes more weight when the world around them is changing faster, as an ideal learner would [115]. A setting that is tuned like that is the goal of a slower loop, so nothing new in kind has been added.

9.1 Plasticity: what it buys and what it costs

Take the simplest learner there is. It holds one value: the right timing, the right weight to expect, the right thing to want. Each experience gives it a noisy sample of what actually fits, and it moves a fraction of the way toward the sample. That fraction is its plasticity. Compare a rut, which moves 5% of the way each time, with a plastic learner, which moves 50% (Figure 11).

In a steady world, plasticity costs precision. When nothing is changing, the rut's typical error is 0.16 and the plastic learner's is 0.58, 3.6 times as large. The plastic learner goes on chasing noise after there is nothing left to learn.

In a changing world, the rut costs more. When what fits drifts quickly, the rut's typical error is 1.61 and the plastic learner's is 0.82. The best setting rises with how fast the world changes compared with how noisy experience is [116]. No setting is best in general.

After a sudden change, the plastic learner is there first. It is within a tenth of the change after 4 samples. The rut takes 45.

A window closes on what it wrote. Let plasticity be high for the first 20 samples and low from then on. If the world during the window was not the world that follows, the learner is left holding the window's value. It fades with a half-life of 34 samples, and 13% of it is still there after 100. Reopen the window in a setting that fits, and what is left is removed. Reopen it in a setting that does not fit, and that is written in just as surely.

Two findings have the same shape. A network that goes on learning has to keep what it already holds while it takes in what is new, which is the stability–plasticity dilemma [117]. And across people, more variability from one movement to the next, in the aspects a task uses, goes with faster learning of the task [118].

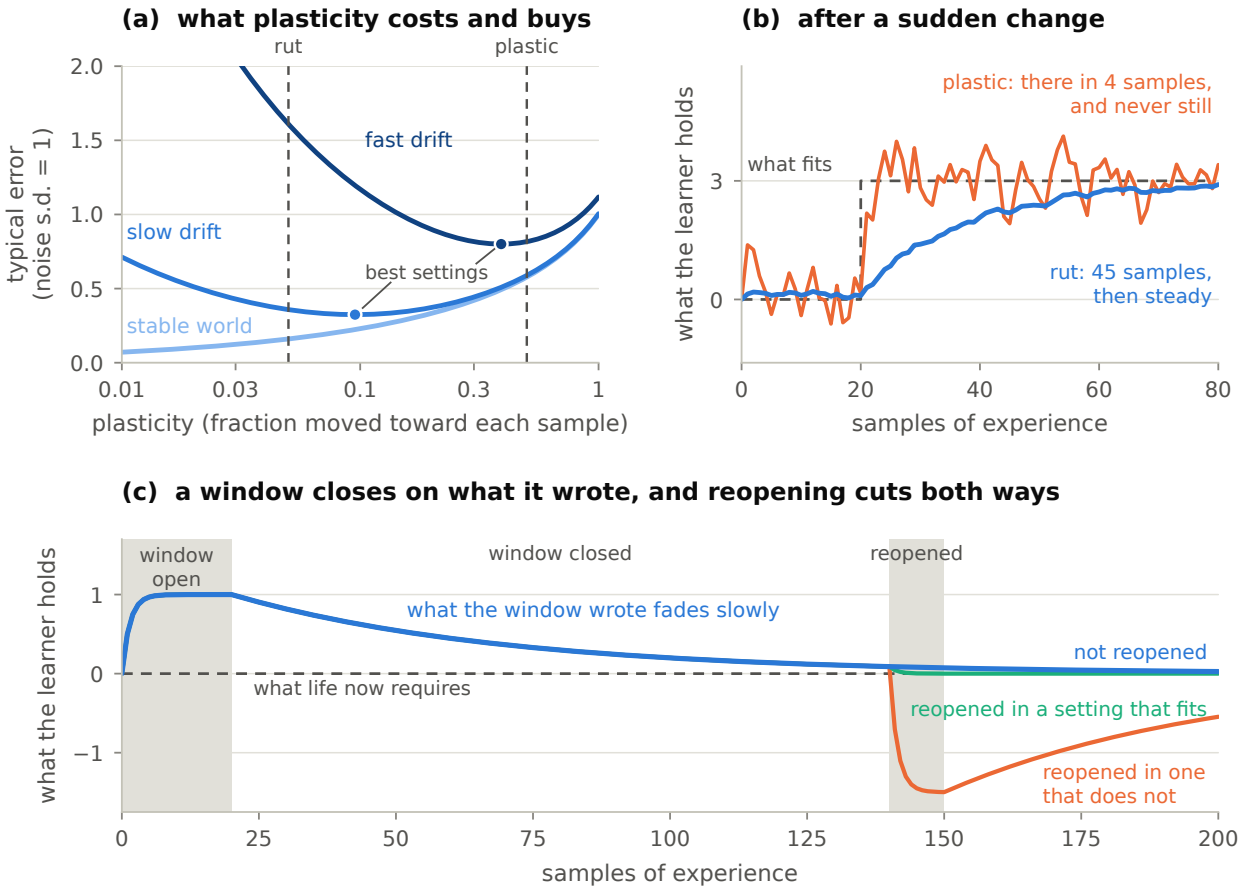

Figure 11: A learner with one plasticity setting. (a) Typical error against plasticity, for a stable world and for two rates of change. Dots mark the best setting, and the dashed lines are the two learners of panel (b). (b) Both learners meet the same sudden change and the same noise. (c) Plasticity is high for the first 20 samples and low afterwards, and during the window the world sits one unit away from where it sits later. What the window wrote fades slowly. A reopened window removes it in a setting that fits, and writes in the new setting where it does not. Panel (c) shows expected values. The equations are in Appendix A.

9.2 What moves plasticity

Childhood. A kitten's visual cortex is most easily changed in its fourth and fifth weeks of life. An eye kept shut for only three or four days in those weeks lost much of its hold on cortical cells, and far fewer cells answered to both eyes. Shutting an eye for more than a year in an adult cat changed nothing that could be detected. And after three months of closure from birth the cortex recovered very little, even when the eye then stayed open for as long as five years [119]. That is panel (c) of Figure 11 in an animal. On one account a sensitive period leaves a circuit committed to a pattern of connections, and experience after it works within that pattern and does not replace it [120]. Children also learn differently. In one study the youngest were the readiest to accept an unusual physical cause that the evidence supported: preschoolers more than school-age children, and those more than adolescents and adults, who tended to stay with a familiar explanation that fit the evidence worse [121].

Reopened windows. A single dose of psilocybin left the dendritic spines of mouse frontal cortex about a tenth larger and more numerous a day later, and the change could still be seen after a

month [122]. Also in mice, a single dose of any of several psychedelic drugs reopened a critical period for learning from social reward, for a time that tracked how long each drug's effects last in people [123]. Ketamine, which blocks NMDA receptors, causes new spine synapses to form in the prefrontal cortex of rats [124]. Two things have to be kept straight. The window is the after-effect and not the drug state: while NMDA receptors are blocked, long-term potentiation cannot be induced, and rats are impaired at learning places [125]. And the evidence is from rodents. For people it is an inference.

9.3 Gain, hold and breadth

Gain. Dopamine has long been tied to the vigor of responding. On one model its steady level reports the average rate of reward, so that more of it makes acting sooner worth the cost [126]. L-DOPA, which raises dopamine function, made volunteers more optimistic, by weakening the updating of beliefs on unwelcome information [127]. In monkeys the responses of dopamine neurons track confidence in a perceptual decision before the choice is made [128]. The same neurons signal errors in the prediction of reward, which is a teaching signal [129], so dopamine touches plasticity as well. In the loop, confidence is trust in the estimate, and trust in the estimate is what licenses gain. Section 3 gave the price: gain makes a loop quick and decisive up to a limit set by its delay, and a system that turns its gain up without shortening its delay acts hard on old news.

Hold. A goal held in mind is a pattern that a circuit keeps alive (Section 5). In biophysical models of circuits that keep one another firing, the activity is stable when it is carried largely by NMDA receptors at the circuit's own synapses [130], so blocking them should loosen what is held. In healthy volunteers ketamine impaired performance on tests sensitive to frontal function, among them vigilance, verbal fluency and card sorting, and it disrupted delayed recall [131]. With less hold, what is done depends less on what was posted a moment ago and more on what is in front of the person.

Sleep loss lowers hold as well. In a pooled analysis of studies of a night or two without sleep, the largest effect was on lapses of attention [132]. After one night, people more often failed to withhold a response to negative stimuli [133]. In the terms of Section 6, new cautions lapse sooner and practiced habits are spared. On that reading a tired person is a less guarded one. Whether that makes for more sociable behavior depends on the study: mice became more social, by a route that depends on dopamine [134], and people became more withdrawn [135].

Breadth. A learner that always takes what looks best never finds out whether something else is better, and one that explores pays for it in the short run [136]. Habits are the low-breadth end. Routines made stereotyped by learning run with little deliberate control [76], and on one account the practiced controller and the deliberate one are each given charge when their own estimate is the more certain [137]. Psychedelics are held to raise breadth from the top. On one model a psychedelic weakens the confidence with which high-level beliefs are held, so that signals from the levels below count for more [138]. In the terms of Figure 7, content that was held so that the world would yield begins to yield itself.

> Hypothesis. **Which setting each state moves.** Dopamine raises gain. Sleep loss lowers hold, of new cautions first. An NMDA blocker lowers hold while it acts. A psychedelic loosens the hold of high-level content while it acts, and raises plasticity for some days afterwards. Childhood has high plasticity and high breadth. Several states move more than one setting,

and the four settings are a way of sorting effects. They are not a claim that the brain has four dials.

9.4 More freedom, or more control?

PROPOSAL. **Range and control trade through plasticity.** With more plasticity a person has more freedom in one sense and less in another. The *range* is wider: more of what the person could come to want and to do is still open. *Control* is weaker, because the precision and skill of the ruts are not there yet or have been loosened. And the danger is greater, because whatever the world happens to present is written in. Childhood has the wide range and the danger. Adulthood has the ruts: a narrower range, held with more skill.

The two senses are the two things this paper has defined. Range is the set of goals a brain can voice, and here the ones it can come to voice. Control is an outcome that follows a held goal through disturbance, and it needs a goal that holds still, a model good enough to serve it, and loops practiced enough to be precise. High plasticity works against all three.

Plasticity also bears on whose the goals are. A system is self-determining to the extent that its setting is a state of the system itself. A plastic system's goals are its own states, and more of the setting of them is done from outside. That is the danger. The closed eye of the kitten is the animal evidence for it. For a reopened window, the evidence is the weight that those who study psychedelics put on context: the drugs' effects are held to depend unusually strongly on it [139], and the authors of the model cited above make any benefit conditional on intention, care and surroundings [138].

There is a safer way to get some of what plasticity gives. A goal can be invented or borrowed and tried on, under a slower loop that keeps it or drops it (Section 7). That is exploring with a brake on. An adult can do it. A young child can do less of it, because the slower loops are among the things still being written.

So the answer is that it depends on the sense. In range there is more freedom in childhood and in any state that opens a window. In control there is less. One setting trades them against each other, and neither end of it is freedom outright. This section describes these states. It does not recommend any of them.

10 What this means for Ring Bank Theory

Ring Bank Theory (RBT) is the author's account of the geometry of conscious experience [16, 85]. This section says what the loop picture means for it. Nothing before this section depends on the theory, and a reader who does not know it can go on to Section 11.

10.1 Two things the theory says, and why both hold

RBT says that *matter cannot act*. No part of the system chooses. The theory describes its own dynamics as relaxation with no agent in it [16, case RB-01].

RBT also says that consciousness ordinarily steers behavior in the near-upcoming time. The Real schema is the "servo target for physical action" [85, §2.2]. The avatar's pose is coupled to the motor system by a servo-link with a gate [16, case RB-02c]. And dissociation is described as the body carrying on with "no conscious servo at the wheel" [16, §1.8], which presupposes that ordinarily there is one.

They do not conflict, for the reason given in Section 1: the first is about what the parts obey, and the second is about what the outcome depends on.

10.2 Two cases that looked like evidence against steering

> Observation. **A night of frames that no one chose.** On the night the theory began (1 January 2022; case RB-01), under heavy sedation, experience broke into discrete frames at about two per second. Throughout the session there was "no experiencer choosing the next frame" [16, case RB-01]. Each frame arrived, and awareness rode the updates without veto. It felt like being run through one's rut responses and simply experiencing one's life. This was not the normal case. The author's reading is that the render of the world model was lagging that night, and the case file records the night's cog-by-cog quality as one that has not recurred.

> Observation. **A ring jerks the knee.** In case RB-02c a ring of attention sweeping down the body passed the knee, and the leg jerked. It did not feel like agency. It felt like a defect: the rings came too close and jerked the knee over. The book's term is proximity-capture of the knee by the access manifold [16, case RB-02c].

Both are exceptions, and the loop says what kind. The night showed two things that have to be kept apart. That no part chooses is always so. That nothing experienced was steering is not always so: that night, experience went on while habit did the steering. Section 3 gives a reason it would. RBT's print, the moment at which content is committed, runs at 7 to 10 a second in the normal state and at about 2 a second under heavy sedation [16, §5.4]. A sample that is held adds half its holding time as delay: 250 ms at 2 a second, against 50 to 70 ms normally. A loop with that much more delay has to give up gain or go unstable. So the slow loop loses its grip, and the habits beneath it carry the behavior. The knee is the other kind of exception. Content moved the body with no goal behind it, which is a fault in the coupling and not an act.

10.3 Where the loop is in the theory

> Proposal. **Where the loop is.** The parts of the loop of Figure 2 correspond to RBT's terms as in Table 7, and they are connected as in Figure 12.

Part of the loop	In RBT	Source
the estimate	the configuration the pinch settles on at the present face: the Real schema R for the world, the avatar in PPS for the body	[16, §1.4, §1.6]
correcting the estimate	pose-force: what the marriage at the access manifold A imposes on a stream that arrives out of agreement	[16, §1.4]
prediction	each stream's orchestrated approach to the next pinch, on the near-future side of the window	[16, §1.6]
the setting	intended content on the near-future side: for the body, the pose the avatar is to take; for an abstract goal, imagined content once it is understood	[16, §1.2]
controller and what acts	the gate and the avatar-to-physical servo-link	[16, RB-02c]
the sampler	the printing operator Π	[16, §1.9, ch. 4]
what sets the settings	understanding, and the aboutness vector upstream	[16, §1.3]
what sounds	the Mic: one high-level voice at a time, driving the print	[16, §1.3, §1.9]

Table 7: A candidate correspondence between the parts of a loop and RBT's terms. Any row could be wrong without touching Sections 1 to 9.

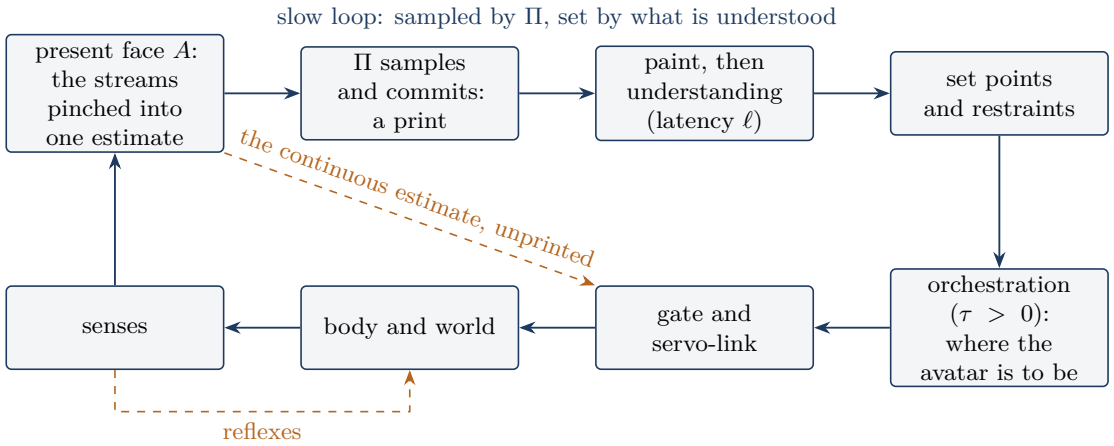

Figure 12: The loop in RBT's terms. Solid: the slow loop, which sets goals and cautions and is sampled by the print. Dashed: fast paths that do not wait for a print. The present face has no control input and needs none. Its output is the estimate, and everything else in the loop runs on it.

An earlier formal paper of RBT read the update at the present face as a filter with no actuator, with action living "upstream" [140]. That reading of the update step stands. What it might suggest about the whole system does not. The upstream is the other half of a loop, and the loop's input is the estimate the present face settles on.

10.4 The print is the slow loop's sample

RBT's print is the moment at which content is committed, several times a second. Which loop is it part of?

> Proposal. **The print belongs to the slow loop.** The print is not the servo. It is the sampled input of the slow loop that sets goals and cautions. The fast loops, reflexes and the corrections of a movement under way, run on the continuous process and do not wait for a print.

A loop sampled seven to ten times a second is too slow for the fast work. Each sample would be 50 to 70 ms stale on average before any other delay is counted. And the fast loops plainly do not wait for awareness. A reach is adjusted to a target that jumps during an eye movement without the jump being perceived [141]. When the jump is perceived, the hand's correction begins more than 300 ms before the person signals awareness of it [142]. Shifts of the seen hand are corrected within about 160 ms and go unnoticed [24].

Goal-setting is the job that can live with a slow sample. Engineers studying human control arrived at the same division: continuous observation, with action updated at intervals [41]. And the rate is right. Perceptual sensitivity and attention fluctuate rhythmically, attention sampling at about 3 to 8 a second and sensory sampling near 10 [143, 144].

> Observation. **A model caught correcting itself in steps.** On opening the eyes in a moving car the road appeared rotated upward, and the access manifold took four ringframes at a print rate of 3 a second to align. The book's card for the case notes: "meta-self saw it instantly; print system caught up ∼1 s later" [16, ch. 3, card "Gradient Descent (Road)"].

10.5 Who steers?

If experience is the estimate that the slow loop runs on, then the question of who steers has a short answer.

> Proposal. **Whoever gets it steers by it.** Whatever grasps the timing, the frames and the content of experience is the same capacity that steers by them, mapping what the scene affords to high-level actions.

The reason is the wiggle test. If changing the content changes the goal that gets set, and changing the neurons beneath the content while keeping the content does not, then the content is what steers, and whatever holds the content is what is steering. Dissociation is the hard case.

> Observation. **Dissociation.** In rare episodes lasting 5 to 60 s the conscious frame comes away from the Real schema. The access manifold, the Bank and the body frame slip together, behind or above the body, and the body goes on acting. It has walked roof joists and answered arithmetic, and it has played bass in time with a song. The book marks the state as a loss of servo lock, with the body carrying on "with no conscious servo at the wheel" [16, §1.8, ch. 3]. "I" was not steering.

> Proposal. **What dissociation separates.** During these episodes something in the brain got the joists, the question and the beat, and steered by them. What it got was not experienced by the "I" that reports afterwards and was not laid down in memory. Getting and steering stayed together. The pair came apart from the self that narrates. Either there were, for that time, two that got and steered, of which "I" was one, or there was one, running the body, and "I" was not part of it. In neither case is there steering without getting.

A body walking joists is a closed loop. It is not an automaton running blind. The book's own phrase for the second alternative is that in dissociation "the lights stay on; the narrator steps out" [16, §1.8]. There is a physiological lead. In mice, dissociative doses of ketamine produce a rhythm of 1 to 3 a second in one layer of retrosplenial cortex, and in one patient with epilepsy a similar

rhythm in deep posteromedial cortex accompanied self-reported dissociation [145].

What in a brain could be the thing that steers? It will not be single spikes. The firing of a cortical neuron varies greatly from one occasion to the next, and quantities appear to be carried by ensembles [11]. A population state is the kind of variable that could pass the wiggle test, and such states are measured: in motor cortex the population state during a reach traces a rotation [146].

10.6 What changes in the theory

- Matter-cannot-act stays. No part of the system chooses, and the present face has no control input.
- The servo vocabulary becomes literal. The estimate is what the pinch settles on, the setting is intended content on the near-future side of the window, and the gate and the servo-link are the controller's path to the body.
- Π is a sampler in the engineer's sense, on the slow loop.
- "Autopilot" means a closed loop that the narrating self is not part of. It does not mean no loop.
- One voice at a time and several goals at once are compatible. The first is about what sounds, and the second about what is in force.
- The question of the level at which experience does work has a test.

10.7 Nine predictions

Table 8 lists what should be found if this paper is right, and what would count against it. None has been tested. Print rate can be varied by sedation and can be tracked from outside, provisionally, by the phase of the perceptual rhythms cited above.

	If the paper is right	It is wrong if
1	When an event calls for a change of goal, how soon behavior shows it depends on where in the print cycle the event fell. Reflexes show no such dependence. A model set right after a sudden reorientation takes about four prints, at any print rate.	goal changes are as indifferent to the cycle as reflexes, or realignment takes the same time at 2 a second as at 10
2	The lead of a moving object over a flash stays the same when the print slows.	the lead grows by about half the print period
3	Where what is experienced and what is on the retina can be set apart, the choice of goal follows the experience, and fast corrections follow the retina, in the same trial.	goal choice tracks the unreported stimulus
4	A caution about the order of two movements shifts one start time and leaves both movements' shapes alone. The margin kept is two to three times the person's timing variability.	caution slows or reshapes the whole movement, or the margin is unrelated to timing variability
5	The careful timing after a flash image fades over seconds, faster under a second task, and a reminder restores it in full. Imagery alone does not change the habit. A real error does.	caution lasts minutes without rehearsal, or the image alone changes the habit
6	A person holds about four new cautions at once. One more displaces an earlier one, which fails all at once. Practiced cautions do not count against the limit.	there is no limit near four, or cautions degrade gradually together
7	During a dissociative episode, behavior stays sensitive to events the person cannot report afterwards.	behavior has stopped responding to what is happening, or everything it responded to can be reported
8	In one person and one task, a condition that speeds re-learning after a change also makes what is held vary more from trial to trial once it has settled.	re-learning is faster with no loss of settled precision
9	After a night without sleep, a newly posted caution lapses sooner, and the timing of a practiced habit is unchanged.	the loss falls on practiced timing as much as on held cautions

Table 8: Nine predictions.

11 What this paper does not show

How a person's brain checks a goal it has made up. For a fly's heading, a mouse's planned lick and a thirsty mouse's drive, the cells are known. For a goal a person thinks up, the step that compares it with how things stand has not been shown, and Section 5 offers a proposal for it and no more. The studies cited were done in flies, fish, mice, rats and monkeys, and we read their abstracts and not their full texts.

That the models are the brain. The five models are the simplest systems that show their points: a one-line loop, an arm with three joints, two movements with one free start time, a caution that lapses at a constant rate, and a learner that holds one number. None is fitted to data. Each shows that an effect exists and which way it goes. The numbers attached to them are illustrations.

That the theorems apply exactly. The stability limit of Section 3 is exact for the loop it describes, and the internal model principle is proved for linear systems. A brain is neither. We use the theorems as exact statements of ideas whose general form is plausible.

That the Observations are confirmed. They are one observer's reports, several made under sedation. We treat them as data about the structure and timing of experience. They motivate the proposals. They do not confirm them, and nothing in the running text that is cited as established rests on them. The predictions of Section 10 are stated so that a second observer, or an experiment with no observer's report in it, can check them.

Why any of it is felt, or whether it is free will. Section 8 says how far the argument goes on each. It gives self-determination and a reason a conscious model would be useful. It stops there.

12 Conclusion

Steering does not need an exception to physics, and it does not need a chooser. It needs a loop. A brain has loops of several kinds: reflexes that balance with no goal stored anywhere, need cells that fire until relief is predicted, and goals held as patterns of activity, which in the fly can be followed from cell to cell. A person's freedom is as wide as the goals their brain can hold and invent. How a brain checks a goal it has made up against the present is the step still missing.

What a person has is self-determination. It is not authorship of oneself from nothing, and whether it deserves the name free will is a question about the name. For Ring Bank Theory it means that matter-cannot-act and conscious steering are both true. Nine predictions say where this could be wrong.

A The models

Every number in the paper is printed by `models.py`, which is distributed with it, and `check_numbers.py` recomputes each one and confirms that the text matches. All five models are deterministic. Where a model has noise, the draws come from a seeded generator.

A.1 Ten runs (Figure 1)

Run i has a state x_i between 0 and 1. It starts at $x_i(0) = 0.08 + 0.84\,(i-1)/9$ and takes 300 steps of $\Delta t = 0.01$:

$$x \leftarrow \min\bigl(1, \max\bigl(0,\ x + \bigl(cK(r-x) + b_i\bigr)\Delta t + \eta_{i,k}\bigr)\bigr),$$

with $c = 1$ when the loop is on and $c = 0$ when it is off, $K = 10$, and the set point $r = 0.5$, or 0.8 in panel (d). The drift b_i is drawn once per run, uniformly within ± 0.06, and the push $\eta_{i,k}$ at each step, uniformly within ± 0.016. The draws come from a generator seeded by the run number, so each run's history of disturbance is fixed in advance. "New pushes" means a different seed.

	Loop off	**Loop on**	**On, new pushes**	**On, set point 0.8**
spread of endings	0.947	0.038	0.059	0.038
mean ending	0.453	0.502	0.490	0.802

Table 9: Ten runs. With the loop on, raising the set point by 0.05 raises the mean ending by 0.0500, and starting every run 0.05 higher changes no ending. Replacing the pushes moves the endings by 0.021 on average with the loop on, and by 0.221 with it off. The weight left on the starting point after 300 steps is $0.9^{300} = 1.9 \times 10^{-14}$.

A.2 The delayed loop (Figure 3)

The loop is

$$\frac{\mathrm{d}x}{\mathrm{d}t} = K\,(r - \hat{x}(t)) + w(t), \qquad \hat{x}(t) = x(t-\delta) \quad \text{or} \quad \hat{x}(t) = x(t-\delta) + \int_{t-\delta}^{t} u(s)\,\mathrm{d}s,$$

where $u = K(r - \hat{x})$ is the push. The first form of $\hat{x}$ is plain delayed feedback. The second adds a forward model: what has been commanded since the senses last reported. The loop is integrated in steps of 2 ms from $x = 0$, with $K = 8\,\mathrm{s}^{-1}$, $r = 1$, and a bump $w = -2.5\,\mathrm{s}^{-1}$ for $1.6 \le t < 1.8$ s.

Without the forward model the loop's behavior is set by the roots of $s + Ke^{-s\delta} = 0$. A root on the imaginary axis, $s = i\omega$, requires $\omega = K$ and $\omega\delta = \pi/2$, which gives the stability limit $K\delta < \pi/2$. A real root $s = -a$ requires $a\,e^{-a\delta} = K$, whose left side is largest at $a = 1/\delta$, so the response stops ringing when $K\delta \le 1/e$. A sample held for h seconds has transfer function $(1 - e^{-sh})/(sh)$, which is $e^{-sh/2}$ to first order: a delay of $h/2$. For a loop with a delay of 100 ms, a hold of 100 ms lowers the largest stable gain $\pi/(2\delta)$ from 15.7 to $10.5\,\mathrm{s}^{-1}$.

A.3 The arm (Figure 5)

With links of unit length, the hand is at $Y(q) = \sum_{k=1}^{3}(\cos\theta_k, \sin\theta_k)$, where $\theta_k = q_1 + \cdots + q_k$ is the absolute angle of link k. The target is $y^* = (1, 2)$. Each step sets

$$q \leftarrow q + \alpha\, DY^{\top} e, \qquad e = y^* - Y(q), \qquad \alpha = \frac{e^{\top} DY\, DY^{\top} e}{\|DY\, DY^{\top} e\|^2},$$

which is the Jacobian-transpose method of robotics [147], and stops when $\|e\| < 10^{-4}$. The factor $DY^\top$ is the model inside the loop of Section 3: it says which joint moves the hand which way. The curve of Figure 5b is computed exactly, by sweeping the angle of the last link and solving the remaining two-link problem for both signs of the elbow.

A.4 Finger and jaw (Figure 9a)

With $\mathrm{sm}(u) = 3u^2 - 2u^3$ on $[0, 1]$, the finger's position is $f(t) = \mathrm{sm}(t/300\,\mathrm{ms})$, from 0 between the teeth to 1 fully out, and the jaw's opening is $j(t) = 1 - \mathrm{sm}((t - d)/200\,\mathrm{ms})$, from 1 open to 0 shut, where d is the lag between the two starts. The finger is clear when $f > 0.6$, at 170.1 ms. The teeth meet when $j < 0.3$, which is 127.3 ms after the jaw starts. The margin is therefore $d - 42.8$ ms. If the lag varies from bite to bite with standard deviation σ_d, the risk of a bite is $\mathrm{Q}(\mathrm{margin}/\sigma_d)$, where $\mathrm{Q}(z)$ is the chance that a standard normal variable exceeds z.

Jaw lag d	**Margin**	**Risk of a bite at $\sigma_d = 25$ ms**
0	−42.8 ms	0.96
100 ms (habit)	+57.2 ms	0.011
150 ms (careful)	+107.2 ms	9.0×10^{-6}

Table 10: The margin and the risk for three lags.

A.5 A caution in volatile memory (Figure 9b, c)

After a flash at time t_f the caution is held until a lapse, and lapses arrive at rate $1/t_w$, so the chance that it is still held is $P = e^{-(t-t_f)/t_w}$. The delay between the flash and the caution's coming into force is taken as zero. A later flash sets P back to 1. While the caution is held the lag is 150 ms, and after a lapse it is 100 ms, so the risk per mouthful is $P \cdot 9.0 \times 10^{-6} + (1 - P) \cdot 0.011$. The figure uses $t_w = 3$ s and $t_w = 1$ s, with flashes at 0 and 14 s.

A.6 Nested loops (Section 3)

Let an outer loop ask for a speed in proportion to a position error, $v^* = K_o(r - z)$, and an inner loop deliver it, $\mathrm{d}v/\mathrm{d}t = K_i(v^* - v)$, with $\mathrm{d}z/\mathrm{d}t = v$. Then

$$\frac{\mathrm{d}^2 z}{\mathrm{d}t^2} + K_i \frac{\mathrm{d}z}{\mathrm{d}t} + K_i K_o\, z = K_i K_o\, r,$$

whose roots are real, so that z approaches r without overshoot, exactly when $K_i \geq 4K_o$. At $K_i = K_o$ the overshoot is 16%, and at $K_i = 2K_o$ it is 4%.

A.7 A learner with one plasticity setting (Figure 11)

Let g_n be the value that fits the world at the nth experience and m_n the value the learner holds. Each experience gives a noisy sample s_n, the learner moves a fraction κ of the way toward it, and the world drifts:

$$s_n = g_n + \epsilon_n, \qquad m_{n+1} = m_n + \kappa\,(s_n - m_n), \qquad g_{n+1} = g_n + \xi_n.$$

The noise ϵ_n has variance σ_s^2 and the drift ξ_n has variance σ_g^2. Write $e_n = m_n - g_n$. Then $e_{n+1} = (1 - \kappa)\, e_n + \kappa\, \epsilon_n - \xi_n$, and for $0 < \kappa < 2$ the variance of e_n settles at

$$E(\kappa) = \frac{\sigma_g^2 + \kappa^2 \sigma_s^2}{\kappa\,(2 - \kappa)}.$$

Setting $\mathrm{d}E/\mathrm{d}\kappa = 0$ gives the best setting,

$$\kappa^* = \tfrac{1}{2}\left(-\nu + \sqrt{\nu^2 + 4\nu}\right), \qquad \nu = \sigma_g^2/\sigma_s^2,$$

which is the steady-state gain of the Kalman filter for this model [116]. A simulation of 200,000 samples gives a mean squared distance of 2.61 where the formula gives 2.59 ($\kappa = 0.05$, $\sigma_g^2 = 0.25$), and 0.333 where it gives 0.333 ($\kappa = 0.5$, no drift).

World	**Rut,** $\kappa = 0.05$	**Plastic,** $\kappa = 0.5$	**Best** κ**, and its error**
steady, $\sigma_g^2 = 0$	0.160	0.577	both approach 0
slow drift, $\sigma_g^2 = 0.01$	0.358	0.589	0.095, 0.324
fast drift, $\sigma_g^2 = 0.25$	1.609	0.816	0.390, 0.800

Table 11: Typical error $\sqrt{E}$ of the learner, for noise of unit variance.

After a jump in what fits, the expected distance is multiplied by $1 - \kappa$ at each sample, so it is a tenth of the jump after $\ln 0.1 / \ln(1 - \kappa)$ samples: 44.9 for $\kappa = 0.05$ and 3.3 for $\kappa = 0.5$. The text gives the next whole numbers. In Figure 11b what fits jumps from 0 to 3 at sample 20, and both learners receive the same seeded noise. In panel (c) the value that fits is 1 for the first 20 samples and 0 afterwards, and κ is 0.5 for those 20 samples and 0.02 afterwards. With the window closed the offset is multiplied by 0.98 at each sample, which gives the half-life $\ln 0.5 / \ln 0.98 = 34.3$ samples and the share $0.98^{100} = 0.13$ left after 100. A reopened window has $\kappa = 0.5$ for 10 samples from sample 140, with the value that fits at 0 or at -1.5.

B Files

Three files come with the paper. `models.py` holds the five models and prints every number. `make_figures.py` draws the computed figures from the same functions (Table 12). `check_numbers.py` recomputes each quoted number and confirms that the text contains it. All need Python 3 with NumPy and SciPy, and the figures need Matplotlib.

Figure	**Functions in `models.py`**
1	`ten_runs`
3	`delayed_loop`
5	`reach`, `fk`, `solution_set`
8	none (schematic)
9	`margin_ms`, `bite_risk`, `p_held`, `per_bite_risk`
11	`learner_mse`, `best_plasticity`, `learner_run`, `window_trace`

Table 12: Where each computed figure comes from. Figures 2, 4, 6, 7, 10 and 12 are drawn in the LaTeX source.

C The marked statements

Kind	Statement	Section
Definition	steering, and self-determination	1
Definition	standing: a goal that starts behavior, and a restraint	6
Observation	a goal arrives as an image that is understood	4
Observation	a flash image while eating chips	6
Observation	a night of frames that no one chose (not the normal case)	10
Observation	a ring jerks the knee (a defect, not an act)	10
Observation	a model caught correcting itself in steps	10
Observation	dissociation: the body acts and "I" does not steer	10
Proposal	the passenger and the ghost share a premise	2
Proposal	the setting is an abstract content	4
Proposal	a goal must settle before it is pursued	4
Proposal	three ways a brain holds a set point (the sorting of Table 4)	5
Proposal	a made-up goal is checked by perception run backwards	5
Proposal	a goal's card: six entries	6
Proposal	the flash is a prediction, and the caution follows when it is understood	6
Proposal	the caution is held, not installed	6
Proposal	four ways goals share one body	6
Proposal	the range of freedom	7
Proposal	set points can be invented and tried	7
Proposal	the loop has settings	9
Proposal	range and control trade through plasticity	9
Proposal	where the loop is in RBT	10
Proposal	the print belongs to the slow loop	10
Proposal	whoever gets it steers by it	10
Proposal	what dissociation separates	10
Hypothesis	how long a caution lasts, and what would make it last	6
Hypothesis	only a few new cautions at once	6
Hypothesis	which setting each state moves	9
Predictions	nine, each with what would count against it	10

Table 13: Every boxed statement in the paper, by kind. Definitions fix words and make no claim.

References

[1] J. Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York, 2003.

[2] S. Yablo. Mental causation. *The Philosophical Review*, 101(2):245–280, 1992. doi:10.2307/2185535.

[3] C. List and P. Menzies. Nonreductive physicalism and the limits of the exclusion principle. *The Journal of Philosophy*, 106(9):475–502, 2009. doi:10.5840/jphil2009106936.

[4] E. P. Hoel, L. Albantakis, and G. Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795, 2013. doi:10.1073/pnas.1314922110.

[5] W. Lohmiller and J.-J. E. Slotine. On contraction analysis for non-linear systems. *Automatica*, 34(6):683–696, 1998. doi:10.1016/S0005-1098(98)00019-3.

[6] R. Landauer. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3):183–191, 1961. doi:10.1147/rd.53.0183.

[7] F. Beck and J. C. Eccles. Quantum aspects of brain activity and the role of consciousness. *Proceedings of the National Academy of Sciences of the United States of America*, 89(23): 11357–11361, 1992. doi:10.1073/pnas.89.23.11357.

[8] S. Hameroff and R. Penrose. Consciousness in the universe: A review of the 'Orch OR' theory. *Physics of Life Reviews*, 11(1):39–78, 2014. doi:10.1016/j.plrev.2013.08.002.

[9] S. Hameroff. How quantum brain biology can rescue conscious free will. *Frontiers in Integrative Neuroscience*, 6:93, 2012. doi:10.3389/fnint.2012.00093.

[10] G. Ryle. *The Concept of Mind*. Hutchinson's University Library, London, 1949.

[11] M. N. Shadlen and W. T. Newsome. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *The Journal of Neuroscience*, 18(10): 3870–3896, 1998. doi:10.1523/JNEUROSCI.18-10-03870.1998.

[12] R. Kane. *The Significance of Free Will*. Oxford University Press, New York, 1996.

[13] M. Tegmark. Importance of quantum decoherence in brain processes. *Physical Review E*, 61 (4):4194–4206, 2000. doi:10.1103/PhysRevE.61.4194.

[14] S. Hagan, S. R. Hameroff, and J. A. Tuszyński. Quantum computation in brain microtubules: Decoherence and biological feasibility. *Physical Review E*, 65(6):061901, 2002. doi:10.1103/PhysRevE.65.061901.

[15] D. M. Wegner and T. Wheatley. Apparent mental causation: Sources of the experience of will. *American Psychologist*, 54(7):480–492, 1999. doi:10.1037/0003-066X.54.7.480.

[16] Brad Caldwell. *Journal of Ring Bank Phenomenology*. Published by the author, Auburn, AL, first edition, 2026. ISBN 979-8-234-11441-9 (hardcover), 979-8-234-11442-6 (e-book).

[17] K. J. Åström and R. M. Murray. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, Princeton, NJ, 2008.

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

[19] N. Wiener. *Cybernetics: or Control and Communication in the Animal and the Machine*. The Technology Press, Cambridge, MA, 1948.

[20] N. D. Hayes. Roots of the transcendental equation associated with a certain difference-differential equation. *Journal of the London Mathematical Society*, s1-25(3):226–232, 1950. doi:10.1112/jlms/s1-25.3.226.

[21] I. L. Kurtzer. Long-latency reflexes account for limb biomechanics through several supraspinal pathways. *Frontiers in Integrative Neuroscience*, 8:99, 2015. doi:10.3389/fnint.2014.00099.

[22] J. A. Pruszynski and S. H. Scott. Optimal feedback control and the long-latency stretch response. *Experimental Brain Research*, 218(3):341–359, 2012. doi:10.1007/s00221-012-3041-8.

[23] S. H. Scott. A functional taxonomy of bottom-up sensory feedback processing for motor actions. *Trends in Neurosciences*, 39(8):512–526, 2016. doi:10.1016/j.tins.2016.06.001.

[24] J. A. Saunders and D. C. Knill. Humans use continuous visual feedback from the hand to control fast reaching movements. *Experimental Brain Research*, 152(3):341–352, 2003. doi:10.1007/s00221-003-1525-2.

[25] I. D. Loram, H. Gollee, M. Lakie, and P. J. Gawthrop. Human control of an inverted pendulum: is continuous control necessary? Is intermittent control effective? Is intermittent control physiological? *The Journal of Physiology*, 589(2):307–324, 2011. doi:10.1113/jphysiol.2010.194712.

[26] E. von Holst and H. Mittelstaedt. Das Reafferenzprinzip: Wechselwirkungen zwischen Zentralnervensystem und Peripherie. *Naturwissenschaften*, 37(20):464–476, 1950. doi:10.1007/BF00622503.

[27] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan. An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882, 1995. doi:10.1126/science.7569931.

[28] R. C. Miall and D. M. Wolpert. Forward models for physiological motor control. *Neural Networks*, 9(8):1265–1279, 1996. doi:10.1016/S0893-6080(96)00035-4.

[29] O. J. M. Smith. Closer control of loops with dead time. *Chemical Engineering Progress*, 53: 217–219, 1957.

[30] R. S. Johansson and G. Westling. Roles of glabrous skin receptors and sensorimotor memory in automatic control of precision grip when lifting rougher or more slippery objects. *Experimental Brain Research*, 56(3):550–564, 1984. doi:10.1007/BF00237997.

[31] J. R. Flanagan and A. M. Wing. The role of internal models in motion planning and control: evidence from grip force adjustments during movements of hand-held loads. *The Journal of Neuroscience*, 17(4):1519–1528, 1997. doi:10.1523/JNEUROSCI.17-04-01519.1997.

[32] R. C. Miall, D. J. Weir, D. M. Wolpert, and J. F. Stein. Is the cerebellum a Smith predictor? *Journal of Motor Behavior*, 25(3):203–216, 1993. doi:10.1080/00222895.1993.9942050.

[33] A. J. Bastian. Learning to predict the future: the cerebellum adapts feedforward movement control. *Current Opinion in Neurobiology*, 16(6):645–649, 2006. doi:10.1016/j.conb.2006.08.016.

[34] R. Nijhawan. Motion extrapolation in catching. *Nature*, 370(6487):256–257, 1994. doi:10.1038/370256b0.

[35] S. J. Blakemore, S. J. Goodbody, and D. M. Wolpert. Predicting the consequences of our own actions: the role of sensorimotor context estimation. *The Journal of Neuroscience*, 18(18): 7511–7518, 1998. doi:10.1523/JNEUROSCI.18-18-07511.1998.

[36] E. Burdet, R. Osu, D. W. Franklin, T. E. Milner, and M. Kawato. The central nervous system stabilizes unstable dynamics by learning optimal impedance. *Nature*, 414(6862):446–449, 2001. doi:10.1038/35106566.

[37] C. M. Harris and D. M. Wolpert. Signal-dependent noise determines motor planning. *Nature*, 394(6695):780–784, 1998. doi:10.1038/29528.

[38] W. T. Powers. *Behavior: The Control of Perception*. Aldine, Chicago, 1973.

[39] K. J. Åström and B. Wittenmark. *Computer-Controlled Systems: Theory and Design*. Prentice Hall, third edition, 1997.

[40] K. J. W. Craik. Theory of the human operator in control systems. I. The operator as an engineering system. *British Journal of Psychology. General Section*, 38(2):56–61, 1947. doi:10.1111/j.2044-8295.1947.tb01141.x.

[41] P. Gawthrop, I. Loram, M. Lakie, and H. Gollee. Intermittent control: a computational theory of human control. *Biological Cybernetics*, 104(1–2):31–51, 2011. doi:10.1007/s00422-010-0416-4.

[42] B. A. Francis and W. M. Wonham. The internal model principle of control theory. *Automatica*, 12(5):457–465, 1976. doi:10.1016/0005-1098(76)90006-6.

[43] R. C. Conant and W. R. Ashby. Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97, 1970. doi:10.1080/00207727008920220.

[44] N. A. Bernstein. *The Co-ordination and Regulation of Movements*. Pergamon Press, 1967.

[45] J. P. Scholz and G. Schöner. The uncontrolled manifold concept: identifying control variables for a functional task. *Experimental Brain Research*, 126(3):289–306, 1999. doi:10.1007/s002210050738.

[46] E. Todorov and M. I. Jordan. Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5(11):1226–1235, 2002. doi:10.1038/nn963.

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

[48] R. A. Adams, S. Shipp, and K. J. Friston. Predictions not commands: active inference in the motor system. *Brain Structure and Function*, 218(3):611–643, 2013. doi:10.1007/s00429-012-0475-5.

[49] B. Hommel, J. Müsseler, G. Aschersleben, and W. Prinz. The Theory of Event Coding (TEC): a framework for perception and action planning. *Behavioral and Brain Sciences*, 24(5):849–878, 2001. doi:10.1017/S0140525X01000103.

[50] A. Newell and H. A. Simon. *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ, 1972.

[51] C. L. Tan and Z. A. Knight. Regulation of body temperature by the nervous system. *Neuron*, 98(1):31–48, 2018. doi:10.1016/j.neuron.2018.02.022.

[52] A. A. Romanovsky. Thermoregulation: Some concepts have changed. Functional architecture of the thermoregulatory system. *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology*, 292(1):R37–R46, 2007. doi:10.1152/ajpregu.00668.2006.

[53] C. A. Zimmerman, E. L. Huey, J. S. Ahn, L. R. Beutler, C. L. Tan, S. Kosar, L. Bai, Y. Chen, T. V. Corpuz, L. Madisen, H. Zeng, and Z. A. Knight. A gut-to-brain signal of fluid osmolarity controls thirst satiation. *Nature*, 568(7750):98–102, 2019. doi:10.1038/s41586-019-1066-x.

[54] W. E. Allen, M. Z. Chen, N. Pichamoorthy, R. H. Tien, M. Pachitariu, L. Luo, and K. Deisseroth. Thirst regulates motivated behavior through modulation of brainwide neural population dynamics. *Science*, 364(6437):253, 2019. doi:10.1126/science.aav3932.

[55] V. Augustine, S. K. Gokce, S. Lee, B. Wang, T. J. Davidson, F. Reimann, F. Gribble, K. Deisseroth, C. Lois, and Y. Oka. Hierarchical neural architecture underlying thirst regulation. *Nature*, 555(7695):204–209, 2018. doi:10.1038/nature25488.

[56] C. A. Zimmerman, Y.-C. Lin, D. E. Leib, L. Guo, E. L. Huey, G. E. Daly, Y. Chen, and Z. A. Knight. Thirst neurons anticipate the homeostatic consequences of eating and drinking. *Nature*, 537(7622):680–684, 2016. doi:10.1038/nature18950.

[57] Y. Aponte, D. Atasoy, and S. M. Sternson. AGRP neurons are sufficient to orchestrate feeding behavior rapidly and without training. *Nature Neuroscience*, 14(3):351–355, 2011. doi:10.1038/nn.2739.

[58] Y. Chen, Y.-C. Lin, T.-W. Kuo, and Z. A. Knight. Sensory detection of food rapidly modulates arcuate feeding circuits. *Cell*, 160(5):829–841, 2015. doi:10.1016/j.cell.2015.01.033.

[59] J. N. Betley, S. Xu, Z. F. H. Cao, R. Gong, C. J. Magnus, Y. Yu, and S. M. Sternson. Neurons for hunger and thirst transmit a negative-valence teaching signal. *Nature*, 521(7551):180–185, 2015. doi:10.1038/nature14416.

[60] L. R. Beutler, Y. Chen, J. S. Ahn, Y.-C. Lin, R. A. Essner, and Z. A. Knight. Dynamics of gut-brain communication underlying hunger. *Neuron*, 96(2):461–475.e5, 2017. doi:10.1016/j.neuron.2017.09.043.

[61] J. Green, V. Vijayan, P. Mussells Pires, A. Adachi, and G. Maimon. A neural heading estimate is compared with an internal goal to guide oriented navigation. *Nature Neuroscience*, 22(9):1460–1468, 2019. doi:10.1038/s41593-019-0444-x.

[62] S. S. Kim, H. Rouault, S. Druckmann, and V. Jayaraman. Ring attractor dynamics in the Drosophila central brain. *Science*, 356(6340):849–853, 2017. doi:10.1126/science.aal4835.

[63] P. Mussells Pires, L. Zhang, V. Parache, L. F. Abbott, and G. Maimon. Converting an allocentric goal into an egocentric steering signal. *Nature*, 626(8000):808–818, 2024. doi:10.1038/s41586-023-07006-3.

[64] E. A. Westeinde, E. Kellogg, P. M. Dawson, J. Lu, L. Hamburg, B. Midler, S. Druckmann, and R. I. Wilson. Transforming a head direction signal into a goal-oriented steering command. *Nature*, 626(8000):819–826, 2024. doi:10.1038/s41586-024-07039-2.

[65] Z. V. Guo, H. K. Inagaki, K. Daie, S. Druckmann, C. R. Gerfen, and K. Svoboda. Maintenance of persistent activity in a frontal thalamocortical loop. *Nature*, 545(7653):181–186, 2017. doi:10.1038/nature22324.

[66] E. Aksay, G. Gamkrelidze, H. S. Seung, R. Baker, and D. W. Tank. In vivo intracellular recording and perturbation of persistent activity in a neural integrator. *Nature Neuroscience*, 4(2):184–193, 2001. doi:10.1038/84023.

[67] S. Funahashi, C. J. Bruce, and P. S. Goldman-Rakic. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, 61(2):331–349, 1989. doi:10.1152/jn.1989.61.2.331.

[68] M. Lundqvist, J. Rose, P. Herman, S. L. Brincat, T. J. Buschman, and E. K. Miller. Gamma and beta bursts underlie working memory. *Neuron*, 90(1):152–164, 2016. doi:10.1016/j.neuron.2016.02.028.

[69] G. Mongillo, O. Barak, and M. Tsodyks. Synaptic theory of working memory. *Science*, 319 (5869):1543–1546, 2008. doi:10.1126/science.1150769.

[70] M. G. Stokes. 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, 19(7):394–405, 2015. doi:10.1016/j.tics.2015.05.004.

[71] E. K. Miller, M. Lundqvist, and A. M. Bastos. Working memory 2.0. *Neuron*, 100(2):463–475, 2018. doi:10.1016/j.neuron.2018.09.023.

[72] J. D. Wallis, K. C. Anderson, and E. K. Miller. Single neurons in prefrontal cortex encode abstract rules. *Nature*, 411(6840):953–956, 2001. doi:10.1038/35082081.

[73] E. K. Miller and J. D. Cohen. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24:167–202, 2001. doi:10.1146/annurev.neuro.24.1.167.

[74] L. R. Peterson and M. J. Peterson. Short-term retention of individual verbal items. *Journal of Experimental Psychology*, 58:193–198, 1959. doi:10.1037/h0049234.

[75] A. Baddeley. Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, 4(10):829–839, 2003. doi:10.1038/nrn1201.

[76] A. M. Graybiel. Habits, rituals, and the evaluative brain. *Annual Review of Neuroscience*, 31: 359–387, 2008. doi:10.1146/annurev.neuro.29.051605.112851.

[77] B. J. Knowlton, J. A. Mangels, and L. R. Squire. A neostriatal habit learning system in humans. *Science*, 273(5280):1399–1402, 1996. doi:10.1126/science.273.5280.1399.

[78] H. K. Inagaki, L. Fontolan, S. Romani, and K. Svoboda. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature*, 566(7743):212–217, 2019. doi:10.1038/s41586-019-0919-7.

[79] R. Basu, R. Gebauer, T. Herfurth, S. Kolb, Z. Golipour, T. Tchumatchenko, and H. T. Ito. The orbitofrontal cortex maps future navigational goals. *Nature*, 599(7885):449–452, 2021. doi:10.1038/s41586-021-04042-9.

[80] H. K. Inagaki, S. Chen, M. C. Ridder, P. Sah, N. Li, Z. Yang, H. Hasanbegovic, Z. Gao, C. R. Gerfen, and K. Svoboda. A midbrain-thalamus-cortex circuit reorganizes cortical dynamics to initiate movement. *Cell*, 185(6):1065–1081.e23, 2022. doi:10.1016/j.cell.2022.02.006.

[81] M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, and S. Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013. doi:10.1038/nature12160.

[82] M. Lundqvist, S. L. Brincat, J. Rose, M. R. Warden, T. J. Buschman, E. K. Miller, and P. Herman. Working memory control dynamics follow principles of spatial computing. *Nature Communications*, 14(1):1429, 2023. doi:10.1038/s41467-023-36555-4.

[83] B. E. Pfeiffer and D. J. Foster. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497(7447):74–79, 2013. doi:10.1038/nature12112.

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

[85] Brad Caldwell. Ring Bank Theory of conscious semiosis: Access geometry, baseline dynamics, and aboutness modulation. Manuscript, April 2026. An earlier version is at https://theoriesofconsciousness.com/conscious-semiosis.

[86] A. Pnueli. The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science*, pages 46–57, 1977. doi:10.1109/SFCS.1977.32.

[87] M. B. Dwyer, G. S. Avrunin, and J. C. Corbett. Patterns in property specifications for finite-state verification. In *Proceedings of the 21st International Conference on Software Engineering*, pages 411–420, 1999. doi:10.1145/302405.302672.

[88] L. Lamport. Proving the correctness of multiprocess programs. *IEEE Transactions on Software Engineering*, SE-3(2):125–143, 1977. doi:10.1109/TSE.1977.229904.

[89] B. Alpern and F. B. Schneider. Defining liveness. *Information Processing Letters*, 21(4):181–185, 1985. doi:10.1016/0020-0190(85)90056-0.

[90] A. Dardenne, A. van Lamsweerde, and S. Fickas. Goal-directed requirements acquisition. *Science of Computer Programming*, 20(1–2):3–50, 1993. doi:10.1016/0167-6423(93)90021-G.

[91] P. J. Ramadge and W. M. Wonham. Supervisory control of a class of discrete event processes. *SIAM Journal on Control and Optimization*, 25(1):206–230, 1987. doi:10.1137/0325013.

[92] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8):3861–3876, 2017. doi:10.1109/TAC.2016.2638961.

[93] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert. Constrained model predictive control: Stability and optimality. *Automatica*, 36(6):789–814, 2000. doi:10.1016/S0005-1098(99)00214-9.

[94] T. A. Henzinger. The theory of hybrid automata. In *Proceedings of the 11th Annual IEEE Symposium on Logic in Computer Science*, pages 278–292, 1996. doi:10.1109/LICS.1996.561342.

[95] D. A. Norman and T. Shallice. Attention to action: willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, and D. Shapiro, editors, *Consciousness and Self-Regulation: Advances in Research and Theory*, volume 4, pages 1–18. Plenum Press, New York, 1986. doi:10.1007/978-1-4757-0629-1_1.

[96] D. A. Norman. Categorization of action slips. *Psychological Review*, 88(1):1–15, 1981.

[97] J. Reason. *Human Error*. Cambridge University Press, Cambridge, 1990.

[98] J. Duncan, H. Emslie, P. Williams, R. Johnson, and C. Freer. Intelligence and the frontal lobe: the organization of goal-directed behavior. *Cognitive Psychology*, 30(3):257–303, 1996. doi:10.1006/cogp.1996.0008.

[99] K. A. Thoroughman and R. Shadmehr. Learning of action through adaptive combination of motor primitives. *Nature*, 407(6805):742–747, 2000. doi:10.1038/35037588.

[100] R. Shadmehr, M. A. Smith, and J. W. Krakauer. Error correction, sensory prediction, and adaptation in motor control. *Annual Review of Neuroscience*, 33:89–108, 2010. doi:10.1146/annurev-neuro-060909-153135.

[101] Y. Nakamura, H. Hanafusa, and T. Yoshikawa. Task-priority based redundancy control of robot manipulators. *The International Journal of Robotics Research*, 6(2):3–15, 1987. doi:10.1177/027836498700600201.

[102] H. Pashler. Dual-task interference in simple tasks: data and theory. *Psychological Bulletin*, 116(2):220–244, 1994. doi:10.1037/0033-2909.116.2.220.

[103] P. Cisek. Cortical mechanisms of action selection: the affordance competition hypothesis. *Philosophical Transactions of the Royal Society B*, 362(1485):1585–1599, 2007. doi:10.1098/rstb.2007.2054.

[104] N. Cowan. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–114, 2001. doi:10.1017/S0140525X01003922.

[105] W. B. Scoville and B. Milner. Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, 20(1):11–21, 1957. doi:10.1136/jnnp.20.1.11.

[106] G. Strawson. The impossibility of moral responsibility. *Philosophical Studies*, 75(1–2):5–24, 1994. doi:10.1007/BF00989879.

[107] D. C. Dennett. *Elbow Room: The Varieties of Free Will Worth Wanting*. MIT Press, Cambridge, MA, 1984.

[108] J. M. Fischer and M. Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press, Cambridge, 1998.

[109] P. van Inwagen. *An Essay on Free Will*. Clarendon Press, Oxford, 1983.

[110] D. Pereboom. *Living Without Free Will*. Cambridge University Press, Cambridge, 2001.

[111] D. M. MacKay. On the logical indeterminacy of a free choice. *Mind*, 69:31–40, 1960.

[112] D. J. Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219, 1995.

[113] M. S. A. Graziano and T. W. Webb. The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in Psychology*, 6:500, 2015. doi:10.3389/fpsyg.2015.00500.

[114] K. Doya. Metalearning and neuromodulation. *Neural Networks*, 15(4–6):495–506, 2002. doi:10.1016/S0893-6080(02)00044-8.

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

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

[117] W. C. Abraham and A. Robins. Memory retention: The synaptic stability versus plasticity dilemma. *Trends in Neurosciences*, 28(2):73–78, 2005. doi:10.1016/j.tins.2004.12.003.

[118] H. G. Wu, Y. R. Miyamoto, L. N. Gonzalez Castro, B. P. Ölveczky, and M. A. Smith. Temporal structure of motor variability is dynamically regulated and predicts motor learning ability. *Nature Neuroscience*, 17(2):312–321, 2014. doi:10.1038/nn.3616.

[119] D. H. Hubel and T. N. Wiesel. The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *The Journal of Physiology*, 206(2):419–436, 1970. doi:10.1113/jphysiol.1970.sp009022.

[120] E. I. Knudsen. Sensitive periods in the development of the brain and behavior. *Journal of Cognitive Neuroscience*, 16(8):1412–1425, 2004. doi:10.1162/0898929042304796.

[121] A. Gopnik, S. O'Grady, C. G. Lucas, T. L. Griffiths, A. Wente, S. Bridgers, R. Aboody, H. Fung, and R. E. Dahl. Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences of the United States of America*, 114(30):7892–7899, 2017. doi:10.1073/pnas.1700811114.

[122] L.-X. Shao, C. Liao, I. Gregg, P. A. Davoudian, N. K. Savalia, K. Delagarza, and A. C. Kwan. Psilocybin induces rapid and persistent growth of dendritic spines in frontal cortex in vivo. *Neuron*, 109(16):2535–2544.e4, 2021. doi:10.1016/j.neuron.2021.06.008.

[123] R. Nardou, E. Sawyer, Y. J. Song, M. Wilkinson, Y. Padovan-Hernandez, J. L. de Deus, N. Wright, C. Lama, S. Faltin, L. A. Goff, G. L. Stein-O'Brien, and G. Dölen. Psychedelics reopen the social reward learning critical period. *Nature*, 618(7966):790–798, 2023. doi:10.1038/s41586-023-06204-3.

[124] N. Li, B. Lee, R.-J. Liu, M. Banasr, J. M. Dwyer, M. Iwata, X.-Y. Li, G. Aghajanian, and R. S. Duman. mTOR-dependent synapse formation underlies the rapid antidepressant effects of NMDA antagonists. *Science*, 329(5994):959–964, 2010. doi:10.1126/science.1190287.

[125] R. G. Morris, E. Anderson, G. S. Lynch, and M. Baudry. Selective impairment of learning and blockade of long-term potentiation by an N-methyl-D-aspartate receptor antagonist, AP5. *Nature*, 319(6056):774–776, 1986. doi:10.1038/319774a0.

[126] Y. Niv, N. D. Daw, D. Joel, and P. Dayan. Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology*, 191(3):507–520, 2007. doi:10.1007/s00213-006-0502-4.

[127] T. Sharot, M. Guitart-Masip, C. W. Korn, R. Chowdhury, and R. J. Dolan. How dopamine enhances an optimism bias in humans. *Current Biology*, 22(16):1477–1481, 2012. doi:10.1016/j.cub.2012.05.053.

[128] A. Lak, K. Nomoto, M. Keramati, M. Sakagami, and A. Kepecs. Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Current Biology*, 27(6):821–832, 2017. doi:10.1016/j.cub.2017.02.026.

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

[130] X.-J. Wang. Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neurosciences*, 24(8):455–463, 2001. doi:10.1016/S0166-2236(00)01868-3.

[131] J. H. Krystal, L. P. Karper, J. P. Seibyl, G. K. Freeman, R. Delaney, J. D. Bremner, G. R. Heninger, M. B. Bowers, and D. S. Charney. Subanesthetic effects of the noncompetitive NMDA antagonist, ketamine, in humans: Psychotomimetic, perceptual, cognitive, and neuroendocrine responses. *Archives of General Psychiatry*, 51(3):199–214, 1994. doi:10.1001/archpsyc.1994.03950030035004.

[132] J. Lim and D. F. Dinges. A meta-analysis of the impact of short-term sleep deprivation on cognitive variables. *Psychological Bulletin*, 136(3):375–389, 2010. doi:10.1037/a0018883.

[133] C. Anderson and C. R. Platten. Sleep deprivation lowers inhibition and enhances impulsivity to negative stimuli. *Behavioural Brain Research*, 217(2):463–466, 2011. doi:10.1016/j.bbr.2010.09.020.

[134] M. Wu, X. Zhang, S. Feng, S. N. Freda, P. Kumari, V. Dumrongprechachan, and Y. Kozorovitskiy. Dopamine pathways mediating affective state transitions after sleep loss. *Neuron*, 112 (1):141–154.e8, 2024. doi:10.1016/j.neuron.2023.10.002.

[135] E. Ben Simon and M. P. Walker. Sleep loss causes social withdrawal and loneliness. *Nature Communications*, 9(1):3146, 2018. doi:10.1038/s41467-018-05377-0.

[136] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, second edition, 2018.

[137] N. D. Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711, 2005. doi:10.1038/nn1560.

[138] R. L. Carhart-Harris and K. J. Friston. REBUS and the anarchic brain: Toward a unified model of the brain action of psychedelics. *Pharmacological Reviews*, 71(3):316–344, 2019. doi:10.1124/pr.118.017160.

[139] R. L. Carhart-Harris, L. Roseman, E. Haijen, D. Erritzoe, R. Watts, I. Branchi, and M. Kaelen. Psychedelics and the essential importance of context. *Journal of Psychopharmacology*, 32(7): 725–731, 2018. doi:10.1177/0269881118754710.

[140] Brad Caldwell. Ring Bank Theory: A variational closure. Manuscript, June 2026.

[141] M. A. Goodale, D. Pélisson, and C. Prablanc. Large adjustments in visually guided reaching do not depend on vision of the hand or perception of target displacement. *Nature*, 320(6064): 748–750, 1986. doi:10.1038/320748a0.

[142] U. Castiello, Y. Paulignan, and M. Jeannerod. Temporal dissociation of motor responses and subjective awareness: a study in normal subjects. *Brain*, 114(6):2639–2655, 1991. doi:10.1093/brain/114.6.2639.

[143] R. VanRullen. Perceptual cycles. *Trends in Cognitive Sciences*, 20(10):723–735, 2016. doi:10.1016/j.tics.2016.07.006.

[144] I. C. Fiebelkorn and S. Kastner. A rhythmic theory of attention. *Trends in Cognitive Sciences*, 23(2):87–101, 2019. doi:10.1016/j.tics.2018.11.009.

[145] S. Vesuna, I. V. Kauvar, E. Richman, F. Gore, T. Oskotsky, C. Sava-Segal, L. Luo, R. C. Malenka, J. M. Henderson, P. Nuyujukian, J. Parvizi, and K. Deisseroth. Deep posteromedial cortical rhythm in dissociation. *Nature*, 586(7827):87–94, 2020. doi:10.1038/s41586-020-2731-9.

[146] M. M. Churchland, J. P. Cunningham, M. T. Kaufman, J. D. Foster, P. Nuyujukian, S. I. Ryu, and K. V. Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 2012. doi:10.1038/nature11129.

[147] B. Siciliano, L. Sciavicco, L. Villani, and G. Oriolo. *Robotics: Modelling, Planning and Control*. Springer, London, 2009.