

## Agency

**Peter R. Killeen, Stephen Helms Tillery & Felipe Cabrera**

**To cite this article:** Peter R. Killeen, Stephen Helms Tillery & Felipe Cabrera (08 Dec 2024): Agency, The Journal of General Psychology, DOI: [10.1080/00221309.2024.2433277](https://doi.org/10.1080/00221309.2024.2433277)

**To link to this article:** <https://doi.org/10.1080/00221309.2024.2433277>



Published online: 08 Dec 2024.



Submit your article to this journal [↗](#)



Article views: 14






View related articles [↗](#)



View Crossmark data [↗](#)



## Agency

Peter R. Killeen<sup>a</sup> , Stephen Helms Tillery<sup>a</sup> , and Felipe Cabrera<sup>b</sup> 

<sup>a</sup>Arizona State University; <sup>b</sup>Universidad de Guadalajara

### ABSTRACT

Agency is action aimed at goals selected by an agent. A deterministic world view leaves scant room for agency. To reconcile the arguments, we represent action as nested control systems, ranging from clearly deterministic to clearly volitional. Negative feedback minimizes deviations from setpoints (goals). Goals are determined by higher modules in a hierarchy of systems, ranging from gamma-efferent spindles through reflexes to operant responses; these last, and the larger system that contains them, called the Self, comprise volitional agents. When operants become habitual they descend to closed teleonomic systems—automaticity. Change in emotional states, and unpredicted changes in the context—raise them back to full volitional systems. At the highest level is the Self—the brain's model of the agent. When aroused out of open teleonomic functioning, it must reconsider means and ends. It does so by simulating action plans, using the same neural systems it uses to effect them. The simulated stimuli and responses are conscious, approximating their perceptions as experienced in real time; this verisimilitude gives them their hedonic value. Positive feedback plays a key role in these complex adaptive systems, as it focuses and holds attention on the most salient percepts and goals, permitting the self-organization of action plans. The Self is not a separate entity, but a colloquy of command modules wearing the avatar of the agent. This system is put into correspondence with Grossberg's Adaptive Resonance Theory. Free will and determinism emerge not as binary opposites, but the modulating inputs to a spectrum of systems.

### ARTICLE HISTORY

Received 4 July 2024

Accepted 18 November 2024

### KEYWORDS

Agents; actions; control systems; consciousness; determinism; free will; intention; simulation; teleology; teleonomy

The purpose of this paper is to provide a scientifically plausible theory of behavior in which both determinism and freedom find comfortable homes. It begins with a basis in classic learning theory, with its many instances of both highly determined behavior and volitional action. It ensconces these in the framework of control systems, where negative feedback modules play a key role in minimizing deviations from setpoints—goals. Issues concerning teleology are addressed, as are those of causality, “downward causation,” and freedom. The “systems of systems” framework is drafted to

help keep the various parts and functions of these control systems organized. Most of the systems are stacked hierarchically, with levels above setting the goals of those below; but the system of these systems is a heterarchy: There are many stacks of percepts and action plans, often competing for dominance. Habitual behavior is the norm, but when a plan or percept is thwarted, reassessment of the heterarchy is required. Conscious appraisal is enlisted to jump out of the loop of that schema, sample the alternate percepts and plans, and coordinate a new stack. This framework is aligned with select neurocomputational theories of cognition. Most of our habitual perceptions and actions are determined by their historical and current context. Once out of that loop, however, our next action, involving positive feedback of attention sensitive to the smallest alteration in its context, is often indeterminate. Our personal stories then write themselves, each with their unique blends of context, logic, fact, and fantasy.

### Agency analyzed

An *agent* is an entity that acts, or has the power to act; *agency* is the power of choosing an action and effecting it. What is acted on is the *object* of that action. Action is typically organized to achieve some goal or achieve it more efficiently. Some movement is pointless, random with respect to any particular goal. When it has a point, however, that is its *objective*. Consider the vendor in a kiosk who threw a stone onto papers to keep the wind from blowing them away. She was the *agent* of that act; the stone she moved was the *object* of that agency, whose *objective* was to secure the papers. An agent may be in the employ of others who choose the goals; or it may be in the employ of itself, who chooses its goals. In the latter case, it is called a *free agent*. A *puppet* is not an agent, but an object often resembling an agent, manipulated in detail by a master or other force, whether present or historical, pulling strings that activate circuits. The purpose of this paper is to develop a model of agency. Agents make choices, and many of those choices are “freely made” in the common sense of that term. They may owe their ultimate origins to often inscrutable initiating causes in their history or that of their species. Working out how this can be so is the goal of this paper.

Agency is the ability, capacity, or power to *act*—i.e., to do some things rather than other things or nothing at all. Exercising this ability involves choosing between alternatives, and that usually entails having preferences for actions that might achieve one’s goals or achieve them more efficiently. Most animals (including most humans) have a measure of agency when they are not comatose, catatonic, or constrained. Offered a carrot and a

meaty bone, a hungry horse will choose the carrot; a dog the bone; a human might refuse both.

### ***The logic of actions***

Agents are entities that can perform actions. The range of the actions that they can perform determines the degree of their agency. What, exactly, is an *action*? And what kinds of beings have a capacity for it? In the 1960s and 1970s, much subtle philosophical effort was put into the first of these questions (e.g., Anscombe, 2011; and Chilsom; see, e.g., Bogdan, 2013). Most of this work was limited to defining what was called *intentional action*—even though the modifier was often dropped (Davidson, 2001). The implicit assumption of this body of work was that only human beings with a capacity for rational deliberation and a command of language were capable of (intentional) action. This limitation of topic had a motive and a disadvantage. Its usual motive was to discover a criterion that would justify assignments of moral or legal responsibility, a necessary condition of which seemed to be that the behavior in question was intentional—i.e., had an objective of which its agent was conscious (cf. Hocutt, 2017). One disadvantage of this limitation was that it limited agency to linguistically capable humans. For more on this topic, see the commentaries on Foxall (2007), and his rejoinder (2008).

### ***The law of effect***

Actions are a subclass of *causes*. How are actions distinguished from other causes? What are the special *kinds* of effects that those behaviors have that would single them out as actions? Thorndike's (1927) *law of effect* may provide a basis. Actions have effects not just upon the agent's environment, but also frequently on the agent herself: An agent learns from the effects of her actions, which reflect back on her, altering her in ways that predictably change her behavior (Hocutt, 2019). Skinner's version of the law stayed outside the organism, referring to the correlation between a reinforcer and a change in the rate of responses in the same class as those that preceded the reinforcer (cf. a different version of that idea in Cowie & Davison, 2016; Davison, 2017). Thorndike's law, however, included a mediator (Mackinnon, 2011): "responses that produce a satisfying effect in a particular situation become more likely to occur again in that situation, and responses that produce a discomforting effect become less likely to occur again in that situation." Thorndike's is a Hebbian model of what reinforcement does: It strengthens relevant neural connections, which are more lasting than changes in behavior, which are often transient, dependent on context and motivation.

Thorndike's law has several advantages over Skinner's: It accommodates stable-state situations in which behavior is followed by a (nominal) reinforcer, but its rate doesn't change (perhaps the animal is responding as fast as it can or is not motivated for that reward); but neural connections can continue to be strengthened even when response probability is at its ceiling or floor. An organism's state, such as satiation or fear, can change without having to rebadge the "reinforcer" as a non-reinforcer or even punisher, and without having to say that the behavior is now at 0 strength: The "satisfying" and "annoying" states, conditioned by dispositions, modulate the animal's response (Killeen & Jacobs, 2017). This is an intrinsic part of Thorndike's version: consumption of pellets is a reinforcer contingent upon the state of satiation of the organism—a state that is a moderator of responding (Mackinnon, 2011). Another important reason to prefer Thorndike's law is that it came with an operational definition of a satisfier: A satisfier is "a state of affairs that the animal will approach." It is a goal. In the next section, we will identify such goals with the reference level of a control system, returning then to some of Thorndike's other prescient observations.

Like other events, actions are frequently followed by certain results—their characteristic effects, the ends that actions of that sort regularly achieve, or achieve with some probability. It was this fact that led David Hume to define causation as an observer's expectation that was based on observations of correlations (Hume, 1748). Observing the frequency of a succession of events, we expect a similar frequency of succession in the future. Such, said Hume, is the *entire empirical* meaning of causation. Hume's list of factors that cause us to attribute causality—temporal and spatial proximity, and "constancy of conjunction": physical and temporal covariation—are the very ones that underlie operant conditioning (Baum, 1973, 2018; Borgstede & Eggert, 2021; Killeen, 1981; Waldmann et al., 2006). Observation of stochastic adaptation to environmental changes—correlations—is what causes an observer to declare a performance voluntary (Neuringer & Jensen, 2010). These same principles apply to primary circular reactions in development, where infants learn to move by correlating an output (e.g., a motor command to arm musculature) and that output's sensory consequences (e.g., my arm just passed in front of my face) (Piaget, 1952). He and Ogmen (2021) instantiated Piaget's insight by constructing a model for goal-directed reaching using physiologically plausible visual processing modules and arm-control neural networks.

### **Models of actions**

Actions are a species of *causes*, identifiable by their *effects*, brought about by *agents*. They are purposeful, if not always consciously purposive,

behavior. This is consistent with Skinner: “When philosophers speak of intention ... they are almost always speaking of operant behavior” (1989, p. 14); and “The field of operant behavior is the field of intention as well as purpose” (Skinner, 1988, p. 609) and “to speak of the purpose of an act is simply to refer to its characteristic consequence” (p. 558). Zuriff (1975, p. 20) agrees that the distinction between automata and agents is the richer control of the latter by the antecedents and consequents of behavior: “Most automata do not acquire new behaviors the way humans do, they are not sensitive to the consequences of their responses in the way humans are, nor does their behavior come under the control of stimuli as does that of a person.”

Skinner boasts that, rather than concern itself with complex processes, such as planning and goal-setting, “operant behavior ... replaces these current *surrogates* of the history of the individual with the history itself” (1988, p. 609, emphasis added). At least, it points at them generically, as the replacement histories are typically available only hypothetically (Burgos & Killeen, 2019; Foxall, 2008), and seldom suffice to explain current behavior—which Skinner seemed to recognize: “The fact remains that direct observation, no matter how prolonged, tells [an observer] very little about what is going on” (Skinner, 1969, p. 9). And that, of course, is because a very important part of what is going on is going on inside the organism.

What did Skinner mean by *surrogates*? A surrogate is, *inter alia*, a woman who has agreed to carry and bear a child for others. Her pregnancy is a current matter of fact; the causes of her behavior are open to interpretation: being historical they depend on the perspicacity of a historian, if there was one; yet we agree that whatever those causes were, they were somewhere in her environment, whether temporally proximate or remote. They are in all senses less salient, less knowable, however, than her current round belly. A surrogate is also an engineering model, constructed when the event of interest cannot be easily or directly measured, so a model is used instead. The model is recognized to be *ad hoc* and fallible but can provide useful insights. Our core model for purposeful action—agency—is a *control system*. It will replace the historical and hypothetical initiating causes of behavior with current surrogates of them, ones that, while very schematic, may provide useful insights. Such surrogate models will carry us a step forward from hypothetical and unidentifiable histories, to hypothetical identifiable models that are acting in the present.

“Private events... may be called causes, but not initiating causes” (Skinner, 1988, p. 486). With that license, and respecting that caveat, we now develop models of agency as causal structures, ones involving both public and private events, that will validate a modified version of his claim cited in the first paragraph of this section, that “Operant behavior involves

intention as well as purpose” (after Skinner, 1988, p. 609). Like we and Skinner, Ginsburg and Jablonka (2019) recognize the key role of learning in the evolution of both agency and consciousness.

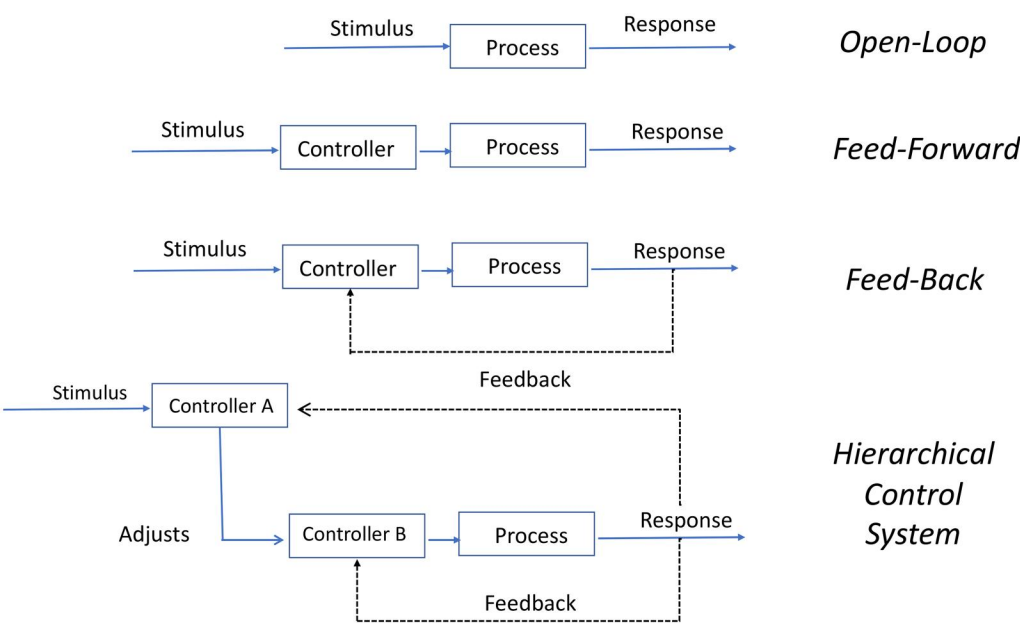
**Control systems**

Norbert Wiener, the inventor of modern feedback systems, and his associates (Rosenblueth et al., 1943) argued that control systems provide the proper models for purposive behavior. A control system is a machine that receives input and generates output. The nature of the input and output depends on the system’s processor, which is programmed by the controller. (For an introduction to control systems in animal behavior, see McFarland, 1971).

**Open-loop systems**

The simplest control system is shown at the top of Figure 1. It is “system” in name only, a process that is the null case of a control system.

**The rock.** Consider the rock tossed onto a pile of papers to stabilize them. The rock moves, but its movement is not an action. The input stimulus is the force exerted by the vendor, causing acceleration of the rock. The response is the subsequent ballistic movement of the rock. Unlike the vendor, the rock is indifferent to where it lands. The rock does not correct



**Figure 1.** Four types of control systems.

course or change in any other way as a function of the proximity to nesting atop the papers. Its movement is an open loop happening, not an action.

### *The puppet.*

Cultural evolution has yielded many types of puppets, from sock puppets through Jim Henson's Muppets to marionettes, figurines controlled by complicated arrangements of strings. Many seem to talk, and some—ventriloquist dummies—seem to converse. They are directly controlled by hands and limbs of their masters. They have no goals, and the effects of their movement change them in no way. Their controller is completely external. The system at the top of [Figure 1](#) is their model.

***Feed-forward systems.*** A more sophisticated control system contains a model of aspects of the environment within the system so that certain changes in the environment trigger appropriate reactions.

***The reflex.*** Many human reflexes are most noticeable in babies; the rooting reflex that finds the nipple, the suckling reflex that obtains milk, the Moro reflex that may have helped an infant to hold onto the mother, the palmar reflex by which fathers delight in lifting their infants by placing a finger in each small palm. Many survive into adulthood (the reflexes, that is): the sneeze, the cough, the startle, the vestibulo-ocular reflex, and many others. Readers of this paper will not forget the yawn reflex, nor we the blushing reflex. The model for reflexes is given in the second diagram in [Figure 1](#). The controller has been shaped by inexorable evolutionary pressures; infants with a good controller survive longer than those with a non-optimal one.

In such open-loop feed-forward control systems there has evolved, either through Darwinian selection or through engineering analysis, a model of some relevant aspect of the world and some typical desirable or undesirable consequence of a response. A sudden sharp pain in one foot while walking signals that we may have stepped on a thorn. If we just reflexively retract that foot we shall fall over, doing worse damage. The crossed-extensor reflex extends the limb on the opposite side to help us maintain balance during the withdrawal of the bad foot. The startle response protects vulnerable parts of our body. These are faster events than actions mediated by learned avoidance responses, short-circuiting the often painful and damaging learning process. Classic unconditioned reflexes and species-specific appetitive and defense reactions (Fanselow, 1997) provide other examples.

On a larger timescale, of course, these are closed-loop systems, ones modulated by the survival of the individual/process that embodies such feedforward models and commands. It is not the refinement of the model



or the act itself that directs behavior: It is the selection among organisms that have the best internalized and integrated crucial aspects of their environment, both external and internal. When a dust or food particle elicits a sneeze or cough response, it is pulling the string to our reflex, and we will clear the appropriate passageway, like it or not. Evolutionary circuits for such reflexes—hardware plans we might call them—may be benevolent masters, may be largely internal to us, yet our public responses are theirs to command. Reflexes are causal sequences of stimulus-processor-response but are not corrigible. It is the ability to refine the model or action that we consider agential, and we cannot do that with such reflexes.

It is important to note two important organizational elements of feed-forward reflexes like sneezing. First, they may operate in part by usurping existing control systems, including some that may be feed-back systems. Breathing is managed by a closed loop reflex system which manages both depth and rates of respiration by dynamically monitoring levels of CO<sub>2</sub> in the bloodstream. Sneezing takes advantage of the same brainstem nuclei and musculature normally used in breathing. Relatedly, they can exist in a complex mesh of control systems which are utilizing the same musculature to strikingly different ends: such as speech *vs.* coughing.

**Priors.** A different type of reflex is perceptual. The world around us has too many degrees of freedom to be resolved unless the brain is willing to make some assumptions about it, assumptions, such as color constancy, kinematics, perspectives, etc. These are most obvious when we see the many illusions in perception texts caused by the miscarriage of such priors.<sup>1</sup> Shepard (1984) speaks of this as “the brain’s extracting the invariants in the environment” and “internalizing those constraints” (also see Shepard, 1981/2017; 1987); and “The rules that govern structures and motions in the physical world may, over evolutionary history, have been incorporated into human perceptual machinery, giving rise to demonstrable correspondences between mental imagery and its physical analogues” (Shepard, 1978, p. 130). There could not be a clearer statement of the configuration of templates for a feed-forward process, one essential for quick disambiguation of the world. Perceptual constancies—invariance in the percepts despite changes in the sensory input (e.g., color constancies) are an important class of perceptual priors (Schulte, 2021). They stabilize our perceptions despite changes in lighting, distance, and other variables that change the physical characteristics of the input arriving at the sensor. Conversely, they can also carry the past into the present to better inform perceptions and dispositions (Honey et al., 2023). Garrigan and Kellman (2008) showed that perceptual learning depends on the constancies, concluding that “constancy-based representations, known to be important for

thought and action, also guide learning and plasticity” (p. 2248). Thus, it is not the incoming sensations, but rather their translation into percepts—representations one level up—by our priors (both innate and learned), that serve as input to higher levels.

**Behavior-altering parasites.** As good as it is to have a playbook written by evolution, any such fixed process can be exploited. There are many gothic examples of animals duping others into feeding them or their progeny. Brood-parasites hijack caregiving for their eggs; cuckoo bees are invited into bumblebee nests by their denizens, where they kill the queen and have their own young nurtured by their new slaves. Caterpillars of the Alcon blue butterfly invite foraging *Myrmica* ant workers to pick them up and take them home and feed them. In most of these cases, the parasites have evolved to exploit releasing stimuli of the host species, who continue to perform their reflexive behaviors most appropriately, only now to someone else’s benefit than that of their genes or their colony. There is no corrective feedback in these feed-forward systems. Cuckolded birds feed a fledgling in their nest that is twice as large as they, without ever pausing to say to themselves “there’s something wrong with this picture.”

A more interesting case arises when parasites not only exploit releasing stimuli but also change settings on control systems in their hosts. The protozoan *Toxoplasma gondii* infects small rodents and causes them to become careless and attracted to the smell of cat urine, which increases their chance of being eaten by a cat, the protozoan’s preferred host. A trematode causes a snail to forage where it is most likely to be acquired by the parasite’s definitive host, waterfowl, during the day; but to move to positions safe from fish at night, as those just eat that parasite, no indigestion. The malaria parasite modulates its mosquito host’s taste for nectar and blood in ways that supports its own survival; if the mosquito is off her diet, no matter. A wasp larva injects its spider host with a hormone that causes it to weave a secure cocoon for the larva to pupate in. There are many other examples of modifying the enslaved host behavior to move the enslaver into ponds or tops of grasses or out into the light to expose them to conditions that favor them; or to transform parts of their bodies to look like the favored prey of their definitive host. They are zombie makers.

These zombies are not puppets. The *Anopheles* mosquito can still fly, and the malaria parasite cannot tell it how to do that. The hosts are largely feed-forward control systems, reflexive animals. What the parasites do is change the hosts’ wiring between the parts of the world that serve as stimuli, and the kind of responses that they make to them. They have flipped switches in the controller of a feed-forward device. Rewiring can induce striking changes in behavior in simple simulated organisms, as elegantly

demonstrated by Braitenberg (1986). These effects may not be limited to insects and rodents: *T. gondii* infects human brains (Parlog et al., 2015), and is associated with behavioral changes and psychopathologies (Tyebji et al., 2019). There is no evidence, at least, that the neuronal rewiring caused by such an infection changes human behavior in ways that are useful to the parasite. Thus, while in rodents, *T. gondii* acts as a feed-forward controller in preferred hosts, in humans it is an open-loop control system, ill-adapted for its human host.

“Mind parasites”—“viral memes”—are metaphorical extensions of such infections. As propaganda they both affect the behavior of their hosts and infect other humans exposed to that host, all to the benefit of an external controller. We have seen too many examples in current affairs to bear mentioning.

**Respondents.** A respondent is “a class of responses defined in terms of stimuli that reliably produce them” (Catania, 2013, p. 462). This includes unconditioned reflexes, set by phylogeny, and also Pavlovian conditional reflexes, set by ontogeny. In the classic literature, they were assumed to be of fixed topography, like the ethologists’ action patterns, and not subject to the effects that they had on the organism. A conditional reflex is one in which the same response made to an unconditional stimulus, such as salivation to the taste of a lemon, will, with repetition, come to be elicited by a novel stimulus—or even by the thought of a lemon. One model for this is the control system in the middle of Figure 1, the feed-forward system. The evolved conditioning process rewires the organism’s model of the world, however, so that now the response is made to a wider range of stimuli. It is a special case of the feed-forward model, one with an ability to make responses conditional on pairs of associated stimuli (or associated stimuli and responses, Donahoe & Vegas, 2004) presented as input. Conditioning generalizes the triggering stimuli, and the nature of those stimuli may affect the nature of the conditioned response (Holland, 1979, 1980; Timberlake, 1994); but the response is not generally modified as a function of its consequences. This is not to say that it is not a functional response; quite the contrary (Domjan, 2005; Hollis, 1997).

Domjan et al. (2000) have explicitly treated Pavlovian reflexes as feed-forward systems. One of the advantages of such a system design, as they and others have noted, is that it reduces response times from those typically available in closed-loop systems. Not only can such lags put the organism at a competitive disadvantage, they can also cause system instability (Dworkin, 1993; McFarland, 1971). Even feed-back systems, however, can develop elegant workarounds so that they can operate as if they were feed-forward models, with limited moment-to-moment dependence on feedback

(McNamee & Wolpert, 2019). Gardner and Gardner (1988) have emphasized the ubiquity of feed-forward in behavioral control, as has Turkkan (1989). Histories of selection and reinforcement condition the controller's priors, but a response, when incited, acts on those, and not on new information to correct the trajectory of the response. Such feed-forward systems can accomplish more than typically realized—but not all that is needed; and sometimes, they miscarry (Breland & Breland, 1961).

**Closed-loop systems.** If the world were perfectly predictable, feed-forward control systems are all that would ever be needed. For organisms living in static environments, such as deep ocean vent-holes, they suffice. Once plants evolved mechanisms to secure water, nutrients, and light, it is unlikely that they would evolve to modify their basic techniques as a function of their success. Evolution has, however, given them taxes, such as heliotropism and gravitropism to achieve their set-points. These constitute some of the simplest of closed loop systems (see Loeb, 1918 for an early general theory of tropisms). They are so named because an indicator of their output is fed back to modify the input, or some other part of the control process, forming an unbroken loop in the system diagram (third panel of Figure 1). Sun and gravity are historically reliable, providing ample opportunity for the evolution of optimal setpoints, and feed-forward mechanisms to satisfy them (Shepard, 1984). In more motile forms, such as animals, the controls are ever-changing. The setpoint may be similar—approach a source of energy as closely as possible—but in the predator, that is very much a moving target. Closed-loop systems alone can adapt to these.

Negative feedback is important because of its generality. The simplest homeostats, as shown in Figure 1, maintain the status quo, with deviations being sent back as instructions to modify the output in the opposite direction, as in thermostats and Watt's steam engine regulator. But any deviation from any set-point can be minimized, and, with hierarchical controllers, those setpoints are themselves continually adjusted. This is goal seeking, with the goal a moving target.

The power of a hierarchy of negative feedback control units such as these was emphasized by Rosenblueth and associates (Rosenblueth et al., 1943), and studied in depth by Powers (1973, 1978). His Perceptual Control Theory holds that action serves perception to minimize deviations from goals; it anticipated a core part of the analysis of this paper. It has successfully passed some experimental tests (e.g., Bourbon, 1996), been generalized to social contexts (e.g., Vancouver & Putka, 2000), and come to be seen by some as a general model for all of psychology (e.g., Marken & Mansell, 2013).

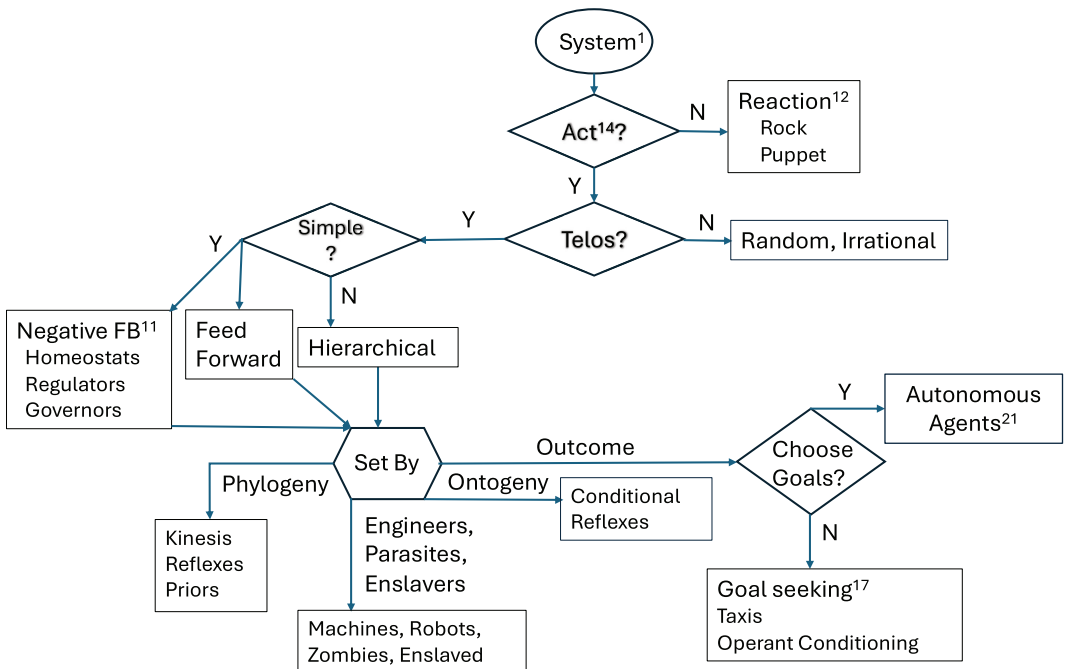
**The operant.** The operant is “a class of responses that is modifiable by the consequences of responses in it. ... Operant behavior has also been called *instrumental* and often corresponds closely to behavior colloquially called *purposive*” (Catania, 2013, p. 453). “When we speak of purposive acts, we mean behavior that is directed toward a goal and is accompanied by a corresponding motivation to attain that goal” (Teitelbaum, 1966, p. 570). To say that behavior is purposeful and goal-directed does not require that its agent be consciously aware of that goal; or even that it be a rational goal to seek (Maugham, 1915). A moth is motivated to fly toward bright lights for reasons that neither we nor it can say. In the above quotes, two behaviorists invite us to consider purposeful—goal oriented—behavior as the central subject of behavior analysis, as Skinner did before them, calling it the *operant*. The systems theorist Ackoff (1971, Sections 13, and 15) makes the useful distinction between *responses*, system events whose antecedents are of interest (viz. respondents), and *behavior*, whose consequents are of interest (viz. operant behavior).

Thorndike’s early analysis of learning is enlightened by the context of control systems. “Animals could not learn to do any act from being put through it, and no association leading to an act could be formed unless there was included in the association an impulse of the animal’s own” (1927, p. 552). “Putting through” an animal’s movements is akin to the first stage in Figures 1 and 2; it does not support learning because the animal is not acting on its environment with feedback (Hocutt, 2019). This conclusion was confirmed in Held and Hein (1963) elegant experiments, and by comparisons between active and passive perception (Gibson, 1962; Prescott et al., 2011). “*Impulse* means the consciousness accompanying a muscular innervation [viz., the signal from the controller in Figure 1] apart from that feeling of the act which comes from seeing oneself move, from feeling one’s body in a different position, etc. [viz., the feedback loops in Figure 1]. It is the direct feeling of the doing as distinguished from the idea of the act done” (Thorndike, 1898, pp. 14–15). The “idea of the act [to be] done” is the *process* in Figure 1. It is a map to a goal. It would wreak havoc if the idea of the act were accepted as a fait accompli for the act. One needs the “direct feeling of the doing” to get conditioned; or punished if a contretemps. This is the only way that behavior-maps can connect with outcome; both in operant and classical conditioning (Donahoe, 2014). Recent reviews have validated Thorndike’s claim of the need for consciousness in conditioning (Lovibond & Shanks, 2002; Skora et al., 2021). This is especially the case where two stimuli are separated in time; consciousness is then necessary to bridge the gap and time-bind them in an association. This resonates with the work of Garrigan and Kellman (2008), who showed that

perceptual learning occurs at the level of conscious perceptions, not at the level of their sensory inputs; of phenomena, not noumena.

### Systems of systems

These diverse considerations call for an organizing framework; living organisms comprise many control systems, from the Krebs cycle which provides the cells' energy, the gamma-efferent system which mediates purposive movement, to the various functional systems that signal the urgency of their goals by fostering motivations—hunger, pain, thirst, cold, to stabilize their *milieu intérieur* (Noble, 2008)—or motivation for better lighting, sound, or haptics, to facilitate achievement of those goals. In modern physiology, the notion of homeostasis has been augmented by the concept of allostasis (e.g., Sterling, 2012), in which systems have evolved or learned to go out of equilibrium to anticipate and correct for more global demands on fitness. Systems based on a chorus of such systems have been studied under the rubric of *systems of systems* (SOS). Ackoff (1971)<sup>2</sup> provides a taxonomy of the field, reflected in Figure 2, one that helps clarify and formalize the above considerations, as we deploy various ensembles of control systems.



**Figure 2.** Behavior treated as control systems. The superscripts index numbered paragraphs in Ackoff (1971).

A *system* is an entity with an interconnected set of elements. An *event* is a change in the structure of the system or its environment. A *reaction* is a system event that is deterministically caused by another event, as in the movements of the rock and the puppet. An *act* is a system event for which no concomitant change in the environment is necessary: The growth of hunger may induce foraging; that of thirst may induce drinking. Acts, therefore, are self-determined events, autonomous changes. Internal changes—in the states of the system’s elements—are both necessary and sufficient to bring about action. “Much of the behavior of humans is of this type” (p. 664). As Skinner would have it, there may be historical events that changed the system so that one particular response is more likely than another in these cases. They are not now required as input; the resulting behavior is, in his words, emitted (like beta decay) not elicited (like a conditioned reflex). Nonetheless, having been emitted, it has consequences which may affect the future likelihood of its emission. If so, it carries us to the bottom of [Figure 2](#).

*Telos* is the end, purpose, or goal of a system; the set-point in a control system. *State-maintaining systems*—most systems with “stats” such as thermostats (the third chart in [Figure 1](#))—may react to internal or external changes. Whereas such systems can be adaptive, “unlike goal-seeking systems, they are not capable of learning because, [un]like Thorndike’s cats, they cannot choose their behavior. They cannot improve with experience” (p. 665). *Goal seeking systems*, on the other hand, ones, such as those found in the bottom charts in [Figures 1](#) and [2](#), can respond differently to an event until a particular outcome occurs. A goal seeking system may be able to achieve the same goal in different ways by engaging different subsystems. If it has a memory, it can increase its efficiency over time in achieving its goal; it can learn. Systems with automatic pilots are goal-seeking.

A *purposeful system* can change its goals: “It selects ends as well as means and thus displays *will*,” as in the rightmost box of [Figure 2](#). How does it do this? Ackoff (1971) defines the *relative value* of an outcome as the relative probability the system will produce/choose that outcome and defines the *preferred* outcome as the outcome that has the highest relative value, which economists call “revealed preference.” Approach to the most preferred outcomes and their signs is a fundamental principle of behavior (e.g., Killeen, 2023). These change, of course, with changes in the state of the organism. “A hierarchy of TOTE units may also represent a hierarchy of values” (Miller et al., 1960, p. 63). To *learn* is to increase efficiency in the attainment of a goal—to reduce the time required to attain it. “Learning can take place only when a system has a choice among alternative courses of action; therefore, only systems that are goal-seeking or



higher can learn.” This extended discussion of Ackoff’s system of systems lays the groundwork for our and Mitchell’s (2023) treatment of agency.

### Teleology

The Greek name *telos* is useful, as it includes a wide variety of things approached, from dry sand above the tidal zone on a beach where a green sea turtle can safely lay her eggs, to a summer vacation at that beach for which we have been saving. Of course, the turtle did not set that goal; evolution set it for her. Teleology—the study of purpose—got a bad rep in biology and in psychology when it was treated as an efficient cause, a trigger for the evolved instinct, or for the operant response (Hocutt, 1974). It is fair, however, to ask what purpose the behavior of the turtle serves, without being accused of having our causal arrows twisted, deeming purpose to be a trigger. The most antiseptic, general, and accurate way to treat *telos* is as the setpoint in a control system. If a response occurs for no goal, neither ones established by phylogeny or ontogeny, and it is not done for its own sake, then it is an irrational waste of energy and time. The teleological branch in Figure 2 carries us back to Figure 1.

Most of Figure 2 concerns behavior controlled by its consequences. In some cases, this sensitivity was set by evolutionary processes, and in other cases, it is updated in real time. Behavior motivated to approach a goal is purposeful. Considering behavior to be under the control of its consequences is not the logical fallacy of *post hoc ergo propter hoc*, mistakenly positing a kind of backward causation (Gould & Vrba, 1982): Efficient causation always works forward in time. Many animals execute such complicated patterns of behavior that it is easy to commit the anthropomorphic fallacy of assuming that they are planning the whole thing with a goal in mind, just as we might. Indeed, too often even we do not understand the causes of our own behavior, but when queried about our reasons, we may point toward rational goals—we “rationalize” it. That typically satisfies the audience, and may even convince ourselves (Nisbett & Wilson, 1977). All of these and other errors have been associated with teleology.

Darwin’s great contribution was showing, in one long argument, that evolution was not *toward* a global point of perfection (cf. De Chardin et al., 1965). This random walk with ratchets, rather than ascension toward perfection, was problematic for Darwin’s Victorian audience. As Dennett (1995) put it, we are often most comfortable with sky-hooks—mythical devices anchored in clouds—that can raise structures from above. But structures that do get built, both by engineers and by evolution, are constructed with cranes, firmly anchored in the ground. Less mystical, but more assured. It often requires small cranes to erect larger cranes—simple



control systems secured, to ground phylogenetic and ontogenetic experiments with higher ordered ones.

Teleology, despite its odor, is thus a difficult concept for biologists and behaviorists to do without (Canfield, 1966). “Teleology is like a mistress to a biologist: He cannot live without her but he’s unwilling to be seen with her in public” (Haldane, quoted in Hull, 1974). Hull went on to note that biologists are now openly wedded to it, and its married name is *teleonomy*. Selection by consequences is intrinsically teleological. Some modern behaviorisms are quite explicitly so (Rachlin, 1992, 1998).

To avoid the sin of philandering with the bad kind of teleology, we must be explicit about the nature of the control by consequences—about the particular feedback system involved—and as clear as possible about the provenance of the control systems that, at higher levels, modulate it. There is a fundamental difference between processes whose control systems lie outside of the process in question, as evolution does to organisms and those that lie within it, as reinforcement does to the responses of individual organisms. The former are puppets and their masters; the latter, agents.

Teleological systems have goals. Feed-forward systems, such as given by the middle diagrams of Figure 1 and the boxes on the left of Figure 2, are goal-seeking: the goal of a frog’s ballistic tongue-flick is to apprehend the fly. Accurate though the movement might be, it is not corrected en route to that goal: It is not corrigible. Evolution set the goal. Mayr (1992, p. 127) calls this level “teleonomic”: “The key word in the definition of teleonomic is *program*.” Mayr goes on: “A program might be defined as *coded or prearranged information that controls a process (or behavior) leading it toward a goal*.... a set of instructions. ... the goal of a teleonomic activity does not lie in the future but is coded in the program” (p. 128. Also see Monod, 1974). These are feedforward systems.

Within the teleonomic category, Mayr notes two subclasses: “Closed” and “Open.” Most such programs, he notes, seem to be closed, refractory to change—hard-core reflexes and instincts. In open<sup>3</sup> teleonomic cases, the event that triggers the instinctive response may be modified over time by experience, motivation, and stimulus context (Timberlake, 1988). Imprinting is an example of this sub-category, as are classically conditioned responses. Many reflexes in higher organisms are of this nature.

These systems stabilize performance with negative feedback. They do this by compensating for perturbations that threaten the achievement of the goal. If a house cools too much, a thermostat will start a furnace, leaving it on until the temperature returns to the set position. It does this by comparing the ambient temperature to a set-point adjusted by the homeowner, thereby correcting the cooling course of the house. A hound following a scent who finds it getting weaker will change course; the direction in which

he is headed may no longer be correct. He will oscillate around his original course until it passes the point of strongest scent, and then the oscillations will be damped as it moves forward.

If teleological control were pure negative feedback, it would quash variation in the output. To make it useful, it must compare that feedback signal to a reference level or template. In servomotors, this could be holding an ideal speed whether under load or off it. A voltage proportional to that speed—or a digital signal of it—is sent to the comparator, which matches it against the ideal and adjusts the process accordingly. Sometimes the controller is adjusted to achieve that desired speed as quickly as possible but without overshoot. In that case, the rate of change of approach to ideal is monitored and the process is adjusted accordingly. This adds differential control to proportional control. Sometimes the process leaves small residual biases; then the integral of that process is taken to eliminate them. These are called PID systems—Proportional-Integral-Derivative. They work so well that 98% of regulatory controllers in refining, chemicals, and paper industry use PID controllers (Åström & Murray, 2010). To work well, they must be tuned to their task and environment. Much of operant conditioning consists of such tuning.

The basic P controller exemplified in thermostats (crude on-off versions of P), centrifugal governors (like the one in Watt's steam engine, a continuous P-controller), and hounds on a scent, are all reactive, not predictive. A basic model of conditioning, the Rescorla Wagner model, is another such error-correcting device (Ghirlanda, 2018; Rescorla, 1987; Rescorla & Wagner, 1972). Regulatory models (Hanson & Timberlake, 1983; Staddon, 2013; Timberlake & Allison, 1974) are more articulated versions of these, as is the delta rule in multi-layered neural net Artificial Intelligence (AI) (Hanson, 1990; Stone, 1986). The standard error-correction technique in AI is back-prop[agation] (Rumelhart et al., 2013), in which each of many controllers have their output weighted as a function of how much they contributed to the pooled error of the last prediction, to improve their contribution to the next. These architectures typically include differential control (Werbos, 1988), and have been used to optimize the performance of classic PID controllers (Scott et al., 1991), and other devices. The modern “deep learning” machines sequence many layers of such controllers. These machines have historically been used primarily as pattern recognizers but can be deployed as pattern generators (necessary in any case to derive a match-mismatch signal in training). The power of such generative AI has become shockingly manifest over the last few years, with the advent of the transformer architectures underlying devices, such as ChatGPT.

Simple drive-reduction models such as Hull's are earlier versions of regulatory systems. A cat chasing a mouse, or a lion chasing an antelope, will,

however, often lead the prey, forecasting where it will be in a few seconds if it continues its trajectory, and aim there. PID systems can do this as well, because the derivative is predicting where the process will be in the next time-step (assuming no dodging). PID control systems take us down to the next level in Figure 2. Sosa and Alcalá (2022) provide an in depth introduction to PID control systems and their relevance to behavior.

Consider next the power steering in a car. Here the reference level is not fixed but varies with the whim of the driver. The driver controls the direction of the car by changing the reference level of the steering wheel. But what controls the driver? Perhaps the desire to return home from work, the instruction to pick up fresh vegetables along the way, the flow of traffic around him, the distance to the car in front, the colors of the traffic light, his hunger, his uncertain memory of the location of the market ... and so on. Some of the actions of driving have become automatized, habitual, descended to the level of open teleonomic reflexes. Others, stopping at the market, the choice of vegetables, the return through the crowded lot to the car, his pulling out into traffic, are voluntary, purposeful, organized by their anticipated consequences.

The driver is a host of many “response guests,” some of which “may be organized by the external environment, or *some of them may organize others of them*” (Baer, 1976, p. 94). The guests’ dialog: “Did I pass the turn-off to the market?” “No, that’s the next turn.” “Do I really need to stop?” “I should.” “Do I think it’s still open?” The environment is recruited: “Yes, my watch says 5:35.” It is these dialogs that give the host agency: The host is Controller A in Figure 1; and at a level above, another controller adjusts those desiderata. This conception of organisms as a hierarchy of control systems is drawn most elaborately by Powers (1973, 1978), and developed as neural models by Grossberg (1980) and Hanson and Negishi (2002). In one of the founding documents of the cognitive revolution, *Plans and the Structure of Behavior* (1960), Miller, Galanter, and Pribram characterized behavior as constituting a hierarchy of control systems, elements of which corresponded to the feedback loops shown in Figure 1, which they called TOTE (Test, Operate, Test, Exit) units. At the highest levels, goals, the set-points of Figures 1 and 2, are not actions or goods, but extended patterns of behavior (Ainslie, 1986; Holland, 1995; Killeen, 2014; Rachlin, 2014); not a single note, but a symphony.

The conversations of Baer’s (1976) “response guests”—of Ainslie’s (1992) “interests,” Ackoff’s (1971) systems, and Cisek’s (2007) “motor affordances”—can become rowdy, with each promoting different goals, or different solution paths to a shared goal. Then “Contradictory reward-getting processes can in effect bargain with each other” (Ainslie, 2005, p. 635), or with their controllers (Selfridge, 1958). One of Hofstadter’s

central themes in *Gödel, Escher, Bach* (1979) was hierarchies of organization of such elements, and the opening of closed teleonomic systems as emergent phenomena. He suggested that “jumping out of the loop” of closed teleonomic feedback systems was accomplished by emotional and motivational shifts. Perturbations of control systems change dispositions, whether benignly, as satiation might, or traumatically, as frustration or fear might. Hofstadter (1982) amplified his, and Baer’s and Ainslie’s, themes of competing interests in his strange book, *I am a Strange Loop* (Adami, 2007; Hofstadter, 2007). Given such competing interests and imperatives, who then, if anyone, “pushes whom around in the population of causal forces that occupy the cranium” (Sperry, 1966, p. 6)?

## Causes

Actions are results, and can also be causes. Their causes may lie in the system itself, as in the growth of hunger; in the environment; in a controller one level up in the hierarchy; or, in times of indecision, in a “heterarchy” (Bechtel, 2019; Cisek, 2007; McCulloch, 1945), with quorum sensing amongst the various interests. Neither causes nor effects are distinctive classes of events with intrinsic features by which they may be identified.<sup>4</sup> As the first philosopher of science, Aristotle, pointed out in his *Posterior Analytics*, the concept of a *cause* (*aition* in Greek) is parasitic on the concept of scientific *explanation*. In Aristotle’s account—a 2500 year anticipation of Hempel and Oppenheim’s hypothetico-deductive “covering law” model—a scientific explanation is an answer to the question “Why”; it is the middle term of a syllogism, the conclusion of which is the explicandum and the premises of which are the explanans (Hocutt, 1974). Thus, consider the question “Why did Socrates die?” A partial answer is “Because he was made to *drink hemlock*.” This statement is elliptical for:

*Those who drink hemlock die;*

*Socrates was made to drink hemlock.*

*Therefore, Socrates died.*

This syllogism illustrates the requirements of a good scientific explanation: a major premise stating a general truth, a minor premise relating this general truth to a specific case via a middle term, and a conclusion stating the effect. The major premise identifies the control system that is invoked, the minor premise the input, and the conclusion the output. In this case (as in most cases), the major premise is only an approximate “general truth”: A more precise description of the control system would specify the amount of the hemlock and the conditions under which it was drunk, and so on. That is OK. Much of science consists of continued

refinement of general laws by boundary conditions and the uncovering of moderators and mediators. The major premise may be construed as a description of the formal cause of an event: The nature of the control system under consideration (Hocutt, 1974; Killeen, 2001). Here, it is a system that is terminated upon the input of a particular drug. The minor premise corresponds to physicists' initial conditions, to logicians' efficient causes, to the input of that control system. Here it is the fatal drink. The conclusion is the output of the control system.

*Why* did the overarching control system make Socrates drink hemlock? Several of Socrates's students and associates had conspired to overthrow Athens's fledgling democracy, and the jury concluded that Socrates and his teachings constituted a dangerous threat to the state. At this level, then, the efficient cause of Socrates's death was making him drink hemlock; the final cause—the purpose of his death—was to eliminate a seditious influence. The goals—the purposes—of a control system are the outputs of those that are one layer above in that system of systems, and those outputs determine the lower level's setpoints.

### **Causal paths**

No event can by itself be either a necessary or sufficient cause. Necessity is instead a “necessary element of a sufficient set of conditions”; if the missing element is inserted, a switch is flipped: A lit match in the context of oxygen and fuel causes a fire. Good (1961) provides a formal treatment of such conditions, and Pearl (2009), critiquing these and other standard treatments, provides a definition of “actual causation” (10.5) and a model for the probability of actual causation given evidence. Causal chains can be extensive, both deep and broad, with the “sufficient set of conditions” themselves embodying causal dependencies with other factors. Whatever sufficiently interesting state of the world you are considering, “you are dealing with a tangled web of cause-effect considerations” (p. 401). For a fascinating introduction to the history of coping with that tangled web, see the *Book of Why* (Pearl & Mackenzie, 2018).

In sum, efficient causes are the input to control systems, whose setpoints constitute the goals or purpose for that unit—their final cause or *telos*—and thereby determine its output. These goals are set by other control systems, perhaps responding to the most strident voices, often modulated by neuro-modulators, such as dopamine, in a pandemonium of such elements. Sperry's question persists: Who pushes whom around, and how, in the population of causes that occupy the cranium? To further the answer to that question requires a close look inside the cranium, and how it might implement something like Pearl's causal path analysis. How such control

systems and causal graphs are embodied constitutes the material cause of these systems, which we examine in the following section.

### **Downward causation?**

Resetting our thermostat has seldom given any of us metaphysical tremors. We hope. Yet it exemplifies “top-down causation.” A partner says, “I’m cold,” and we obligingly turn up the thermostat. That causes electrons to flow to our heat exchanger, which milks the environment for calories. Many physical changes were caused by a casual social exchange, ones that are not explicable at any level other than the psychological. While this is not problematic for most of us, it is for a few. Hard-core determinist (“hard incompatibilist” in modern terminology) will opine that all of my behaviors, and thus those of the machinery it engaged, were determined by constellations of atoms and molecules that caused me to accede, perhaps unwillingly, to my partner’s request. Well, those atoms were there and operative for sure; but did they take into account our recent argument? My preference for a cool environment? What sense does it make to freight those atoms with the causal burden of adjusting a thermostat? The noted cosmologist GFR Ellis thinks that it makes little to none.

These issues, as simplistic as they may seem, are central nonetheless. Understanding emergence and levels of causality is key to understanding the possibility of agency. In one of his many salient articles on emergence and downward causality, Ellis (2012, p. 129) noted five types of top-down causation (*tdc*): (1) algorithmic *tdc*, as when computer code flips the bits in a computer; (2) *tdc* via non-adaptive information control, as in the various types of homeostasis; (3) *tdc* via adaptive selection, from the context to the system, as in Darwinian selection; (4) *tdc* via adaptive information control, as in associative learning and predictive perception; and (5) intelligent *tdc*, as in the construction of a new jumbo jet, or the education of a new human being. The roles and values that are inculcated in the latter case will guide the individual’s actions and goals as adults. Meaning and purpose “form a high level in the hierarchy of causation in the mind” (p. 131). The last three instances are complex adaptive systems, involving interactions of *tdc*, bottom-up control, and control by context. Most of these cases may be mapped onto Figures 1 and 2. The context, with all of its randomness, plays a role, often historical, in all three cases. But complex systems have their own logic and generate their own patterns that seldom can be predicted from lower layers or context; they are in that sense autonomous. *Tdc* is not inconsistent with determinism; it vests it in all of those levels. The lower levels are necessary, but seldom sufficient, causes of action; the top levels, interacting with context and lower levels, are the agents of

action. Ginsburg and Jablonka (2019), channeling Aristotle, identify 3 nested levels of organizational goal directedness: Reproductive (Ellis's Levels 2 and 3); Sensitive (Level 4); and Rational (Level 5). To these authors, it is at Level 4, mediated by associative learning and episodic memory, that consciousness arises.

### ***Complex dynamic systems***

Juarrero (2000) has focused on the salience of complex adaptive systems (*cas*) as key to understanding intentional behavior. She characterizes them as involving:

positive feedback processes in which the product of the process is necessary for the process itself. Second, when parts interact to produce wholes, and the resulting distributed wholes in turn affect the behavior of their parts, inter-level causality is at work. Interactions among certain dynamical processes can create a systems-level organization with new properties that are not the simple sum of the components that constitute the higher level. In turn, the overall dynamics of the emergent distributed system not only determine which parts will be allowed into the system: the global dynamics also regulate and constrain the behavior of the lower-level components. (p. 26)

Juarrero rightly notes the importance of positive feedback in intentional systems. Unlike negative feedback, in which deviations from a goal are damped, with positive feedback deviations toward a goal are amplified.<sup>5</sup> This is true in predictive perception—"Is that a stick in the trail, or a snake?"—where it amplifies focus on the object, and further on features that might discriminate the two hypotheses. It is true in motivation and intention. An evening walk wafts odors from the neighbor's kitchen; we wonder what we shall prepare for our own supper; we review possibilities in the refrigerator, and find we are starting to get hungry. We plan to reheat some of last night's dinner and prepare a fresh salad. It is no longer the sunset, but the salad that has captured our attention, and organized our action plans. Intentions.

a very important feature of self-organizing dynamical systems, their organization itself determines the stimuli to which they will respond. By making its components interdependent, thereby constraining their behavioral variability, the system preserves and enhances its cohesion and integrity, its organization and identity. (Juarrero, 2000, p. 39)

In a few words: positive feedback in self-organizing dynamic systems constitutes them as intentional agents.

### ***Computational neuroscience***

Stephen Grossberg has spent his career in pursuit of a biologically cogent, coherent account of the mechanisms of the brain, arrayed in a constellation



that is in close alignment with many perceptual and motor effects in the literature. This account, absent its important mathematical underpinnings, is summarized in his magnum opus, *Conscious Mind, Resonant Brain* (2021a), while a more detailed compendium of his work through 1980 can be found in Grossberg (2012).

*Gated dipoles*, opponent process arrangements of neurons, are a basic element in Grossberg's neural nets. Configured in on-center off-surround modules, these underlie many sensory phenomena, such as contrast, filling-in, and illusions. Moving up to the level of perception, Grossberg echoes the many observers who have made a distinction between seeing and recognizing. Seeing is largely a bottom-up, feed-forward process; recognizing is an interaction of that and top-down processes. The top-down processes are essentially hypotheses about the contents of the scene—the perceptual priors discussed above. When they misalign with the input, they lead to illusions, “the belief that there is a line or a shape at that position” (Mitchell, 2023, p. 116).<sup>6</sup> When they align, they are locked into a resonance. Shepard (1981/2017) understood the centrality of this metaphor and reviewed its history in the literature on perception and thinking (Shepard, 1984, pp. 433 ff.). It is not just a metaphor, however: “the neural networks in our visual systems are not only figuratively tuned, they are literally harmonic systems (Ratliff, 1983, cited in Shepard, 1984, p. 436). But there is a gap between Shepard's metaphors and Ratliff's data. This gap is bridged by Grossberg's *Adaptive Resonance Theory* (ART).

ART posits a distributed field of excitation from an input, stimulating a field of learned categories (hypotheses about its contents). Whichever brain model comes closest under a lax criterion (modulated by the “vigilance parameter”), is further scrutinized—the vigilance parameter is made more stringent. If it continues to match, the resonance strengthens connections—adaptive weights—between the category and the scene. If the match then becomes faulty, a large mismatch signal resets the search and jumps to the next most likely candidates. To avoid dithering between two equally plausible percepts/plans like Buridan's ass between two equally attractive goals, the circuits that process these decisions have a winner-take-all configuration (Grossberg, 2021a; Mitchell, 2023, p. 124). This causes the system to jump to a new, then the winning, cell population/reckoning (Grossberg, 2021b), not unlike Hofstadter's (1979) “jumping out of the loop.” These ART modules, a sophisticated version of Hebb's “cell-assemblies” (Nadel & Maurer, 2020), exist on many levels, from basic perception to cognitive-emotional resonances (Grossberg, 2021c, Figure 1). They are an embodiment of Dennett's “multiple drafts” model of consciousness (Dennett & Akins, 2008). He and Mitchell note the brain systems where these are likely to occur (e.g., “The tectum, located at the roof of the midbrain, is designed



to map incoming sensory information onto a set of possible actions” Mitchell, 2023, p. 122; “the basal ganglia ... sit between the cortex and thalamus in a nested set of loops ...” p. 125). Ashby and associates (Ashby et al., 2024) provide the most current brain model of agency, and Koch (2004) an alternative perspective. A refinement of ART, LAMINART, takes advantage of the laminar structure of the cortex, giving territory for a hierarchical interaction of the adaptive modules. In Grossberg’s ART and its variants, conversation is never just among guests, interests, or TOTE units. It is among them in the context of an input to be matched and resolved, a motor pattern to be realized. The input could be a sound or sight; or it could be competing drives, such as hunger and fear of predators. The competition is resolved, in light of the input, by the winning categories or motor plans, biased by a history of conditioning in like circumstances. The results of this conversation constitute top-down control, or “downward causation” (Laughlin & Pines, 2000; Littrell, 2008).

### ***Neural reflects of agents in action***

In a normal resting state, a set of medial cortical regions involved in executive function and control settle into coordinated oscillation with other areas of cortex associated with processes like language, spatial processing, and self-awareness. This Default Mode Network (DMN) was initially identified as a set of cortical areas that decreased activity when a subject undertakes a task (“task-negative regions,” see Raichle, 2015 for a review). More recent evidence suggests that this system supports reflection—discursive thinking. For a subject in an MRI tube, only when the screen springs to life does attention focus on the task at hand. Oscillations in the DMN fade to the background and now shift to networks supporting the experimental task. Then, synchrony arises amongst a set of cerebral regions involved in localizing auditory and visual stimuli, along with the sensorimotor cortices around the central sulcus, and the Sensorimotor Network executes the movement. Even simple tasks like moving to a visual target will occasion evolving transitions between multiple networks over time. When the subject’s screen illuminates, the Dorsal Attention Network will take over, seeking resonance between top-down attention and bottom-up sensation to cue upcoming task demands. Once the cues are provided, the Control Network takes over, with its set of attentional and executive modules. The Control Network plans the response, and then hands control over to the Sensorimotor Network. As needs change, the cortical tissue brought to bear changes with it. These actions are the workings of an autopoietic ensemble (Bechtel, 2013; Beer, 2004; Maturana & Varela, 1991).

Carhart-Harris et al. (2014) likened the DMN to the ego: It chooses. Psychedelic drugs disrupt this agency. Among their myriad cognitive and perceptual effects, psychedelic drugs suppress the sense of self, with boundaries between the self and the surrounding world become fluid (Luke & Terhune, 2013). How? These drugs suppress the DMN, which referees amongst the other networks. Grossberg has shown how the active network can be self-selected based on the intersection of vigilance parameters and external inputs. Symptoms like synesthesia could be a result of disruption of the vigilance and selection parameters leading to insolvable competitions between networks, exacerbated by psychedelics. Psychedelic sight is a liberated version of normal sight, with weaker tethers to the input.

## Consciousness

In lines written a century before Skinner's defenestration of the mind, another Harvard professor rebutted the mechanistic theories of Huxley, Balding, and Clifford. In his essay "Are We Automata?" William James first characterizes their position in this seeming caricature:

Feeling is a mere collateral product of our nervous processes, unable to react upon them any more than a shadow reacts on the steps of the traveler whom it accompanies. Inert, uninfluential, a simple passenger in the voyage of life, it is allowed to remain on board, but not to touch the helm or handle the rigging. The theory also maintains that we are in error to suppose that our thoughts awaken each other by inward congruity or rational necessity, that disappointed hopes cause sadness, premises conclusions, &c. The feelings are merely juxtaposed in that order without mutual cohesion, because the nerve-processes to which they severally correspond awaken each other in that order. ... The theory itself is an inevitable consequence of the extension of the notion of reflex action to the higher nerve-centres. (James, 1879, pp. 1–2)

James observes that for a person—especially one with so astute a mind as Huxley's—to assert that he is conscious—but that his formidable mind is but a lazy bystander to his behavior—makes him a dualist. James enlists Huxley's idol Darwin as a foil, asking, given the struggle for survival and its premium on efficiency: "Of what use to a nervous system is a super-added consciousness? Can consciousness increase its efficiency by loading its dice?" (p. 4).

We think so, as James did: Even deep neural nets load their dice through attentional biases (Hanson et al., 2018). James observed that consciousness retreats from tasks where it is no longer needed—it retreats to habits, to teleonomic responses, to the unconscious details of riding a bike or tying a shoe; to conditioned reflexes. Here the conscious "mind" is neither engaged nor necessary: "in the case of all skilled actions, whether tying your shoelaces, playing a musical instrument, or dribbling a basketball—the mind goes

elsewhere while the body performs” (Dretskey, cited in Metzinger, 2013; Sutton et al., 2011, p. 90)—or perhaps to a very primary level of consciousness (Mansell, 2024). But in new, more complex situations where the system must reconsider, at the highest level of control, goals (is that gazelle too far to profitably chase?), action plans (can I cut it off?), danger (will those hyenas overwhelm me?), it is important to jump out of the reflexive loop, to a higher level of consciousness (Mansell, 2024). Some quick simulations of perceptions (is this really what I think I am seeing?) and actions (when should I emerge from cover, and where?) clarify the understanding of which are most hopeful. We call a convention of our response guests for a quick problem-solving mockup and walk-through (Barsalou, 2009; Grush, 2004; Hesslow, 2002). This is what the competitive dynamics of ART achieves. These simulata are not psychic entities, but embodied representations of our historical experiences (Tonneau, 2004), sewn together in novel contexts by the host; the system becomes inseparable from its history of experiences and transformations (Varela et al., 1993). The host—the forces inside the cranium, or more accurately in the nervous system as a whole, often in concert with its environment—finds resonances on many levels and the winning combination at the highest level (of the cortex; of LAMINART) is enacted. The importance of these perceptions, plans, and potential consequences, and their causal, operant (goal-seeking and corrigible) nature, is why Dennett (1975) argues that the Law of Effect will never go away.

Embedded in connectionist machines as *reinforcement learning* (RL), such feedback constitutes a powerful computational technique. The RL modules, which are intrinsically feedback systems that optimize the weights of a non-linear multidimensional curve-fitting machine, can be stacked in layers (Hanson & Burr, 1990; Parr & Russell, 1997) not unlike LAMINART. The addition of an attentional mechanism (Vaswani et al., 2017) was key in designing the most powerful “transformer” architectures of modern AI. The architecture of ART functions in ways like the transformer machines, and has found useful employment in industry (Grossberg, 2021a). Unlike deep learning machines, ART also provides a coherent model of how the brain functions, even providing hints as to the role of consciousness in cognition.

### ***What do we really see?***

Most of us know that the projection of images onto our retina is inverted (No problem: our brains reinvert, or at least reinterpret them, somehow). Fewer of us are familiar with the amazing research, conducted in the late 19th century, that re-inverted those images, turning the world upside down.

George Stratton (1897), like many committed scientists and physicians, performed a heroic experiment on himself. For a week, the only way he saw the world was through a pair of inverting lenses. Slowly, object by object, things re-inverted: On the second day, in his Kafkaesque report: “I could, for instance, voluntarily bring before me, in consistent relation to the visual field, the general outline of the room in which I was sitting. My own body, however, was much less tractable; at best I could get only my legs and arms appropriately represented, and this only by an effort not required by other objects” (Stratton, 1897, p. 347). On the third day:

The representation in the old [natural] way, though, was the spontaneous one, and doubtless was always at least in the background. But in this older representation there was an unusual paling and weakening of the image of those parts which had most often been seen during the course of the experiment. By bringing my legs and arms into view, the older representation became a sort of torso, the filling in of the seen parts refusing to appear, except in the vaguest way, even by an effort of will. When objects other than the body were in sight, they were not (p. 351).

By the fifth day, his seeing had become even more fully conditioned: “At the *thought* of putting on the lenses, in the morning, there was an influx of ideas in the new visual form. I even noticed in many cases that there was a reconstruction, in the new terms, of objects which I had just before been thinking of in the old way” (pp. 354–355; emphasis added).

So, what was Stratton really seeing?

What do we see when looking at the world? “It”, or “a reconstruction, in new terms, of objects which you had before been thinking of in old ways”?

In discussing the evolution of optics on smart phone cameras, an expert noted the importance of such [re]construction in the latest devices:

It’s a shift from, “Let’s make images less noisy and sharper to let’s recreate your memories”, because your memories are different from reality, and that’s okay. That’s a perception and human thing. Let’s build for humans.

When I note the difference between capturing reality and memories, Reynolds defends camera creation versus capture. “Your memories *are* your reality. What’s more real than your memory of it? If I showed you a photo that didn’t match your memory, you’d say it wasn’t real.” (Raymond Wong, discussing phone cameras with Isaac Reynolds, on Inverse, April 2024)

What is real then? Stratton’s “ideas in the new visual form” or “objects thought of in the old way”? What the iPhone captured, or “your memory of it”? *Are* “your memories your reality”?

The degrees of freedom in visual and auditory scenes are overwhelming. We interpret those scenes by matching them to things we think they may be. This is a negotiation between bottom up and top down, with the best top-down candidates winning. “Best” means they account for most of the

relevant degrees of freedom in the input—the part of the input we are coming to focus our attention on. Our reality is neither in the input nor in the memory, but in the module that best resonates between top and bottom. That resonance locks in a percept, continually updated and fine-tuned. Many diverse processing centers in the brain contribute to it. In Seth's (2021) analyses, these are the “controlled hallucinations,” constrained by both sensory input and our templates for perception and action, that constitute our mental life. Our reality is neither the *ding an sich*, nor the memory, but rather the resonance between them, the percept. We do not see things as they are, nor do we see them as we are; we see their resonance in the mind's eye.

If in turn there is a resonant connection of percept with an action system, then that is automatically invoked, and a reflex or a habitual response ensues. In the first days of Stratton's new experience such responses were continually frustrated, requiring continual adjustment, renegotiation, retuning, between sight and action. Maps from sensation to perception were redrawn; habits undone and redone. Progress was tedious. As with perception, so also with action: Bartlett (1995) spoke of returning a tennis volley: “When I make the stroke, I do not produce something absolutely new, and I never merely repeat something old. The stroke is literally manufactured out of the living visual and postural ‘schemata’ of the movement and their interrelations [with the context]” (Bartlett, 1995, pp. 201–202). The symmetry is notable: Of memory and sensation together creating percepts; of schemata/motor plan and context together creating action, of action that is kinetically tailored to the situation, just as percept is to sensation. In all cases, weights are subsequently adjusted.

The advantage of habitual action is that it is quick and effortless: Once learned, response plans—“phase sequences” in Hebb's terminology “schemata” in Bartlett's, “plans” in Miller Galanter and Pribram's, “reflexes” in Pavlov's—can be run off quickly, without much central supervision, solving Lashley's “Problem of serial order in behavior” (Rosenbaum et al., 2007). They flow (Nakamura & Csikszentmihalyi, 2001). This is attested by baseball's great philosopher, Yogi Berra when asked what he thought about while he was on the plate: “Think? How can you think and hit at the same time?” (cited in Sutton et al., 2011, p. 80). Why then think at all? “Thought is very helpful when we are in novel or important circumstances, the rest of the time it rather gets in the way. We only think when our habits give out” (Pollard, in Sutton et al., 2011, p. 88). Or when we play with ideas (Shepard, 2008). Neural models of the conversion of thoughtful to habitual responses are found in Kovacs et al. (2021).

What then is this labor-intensive, time-consuming activity called *thought*, called into play when our habits give out, and what is its value? It occurs

when there is a strong mismatch signal: Something we are seeing, or something we are doing, just isn't right. "System 2 is activated when an event is detected that violates the model of the world that System 1 maintains" (Kahneman, 2011, p. 24). That mismatch jumps us out of the loop of business as usual. We must decide among various alternative perceptual modules, and various action modules. When a decision is made to execute a new plan, "It is taken out of storage and placed in control of the information processing capacity. It is brought into the focus of attention ... Usually the plan will be competing with other plans in the process of execution, and considerable thought may be required ... The parts of a Plan that is being executed have special access to consciousness ... that is necessary for coordinating parts of different Plans" (Miller et al., 1960, p. 65).

How is the decision among percepts and plans made? Alternate potential plans are simulated by enacting the relevant perceptual modules, and the associated relevant action modules. These are the same modules active in perception and action. Ideomotor responses, "'Ghost gestures' (Behnke, 1997) are tendencies toward movement, schematic or barely perceptible ghostly micro-movements that can persist in the body even when the implied or virtual larger-scale gesture or bodily pattern is not actually performed" (Sutton et al., 2011, p. 94; also see Mechner, 1995). Percepts and associated ghost gestures are quickly played out, and their potential success evaluated (Hesslow, 2002). Evolution favors economy of means, and what is more economical (and veridical) than using the same modules for contemplation that we do for perception and actions (Anderson, 2014)? Relevant visual and motor circuits are indeed activated when we think (Anderson, 2003; Killeen & Glenberg, 2010; Nieder et al., 2020; Proffitt, 2006; Shapiro, 2010, 2014; Wilson, 2002); cognition is embodied, using the same control processes for thinking as for seeing and acting. Indeed, Cisek (2007, p. 1585) proposes that "the brain processes sensory information to specify, in parallel, several potential actions that are currently available. These potential actions compete against each other for further processing, while information is collected to bias this competition until a single response is selected," specifying where in the brain the competition takes place, and crafting a computational model of it.

In a series of renowned experiments, Libet (1985) recorded readiness potentials (RPs) from the brain 350 ms before subjects reported a conscious choice to move. "The volitional process is therefore initiated unconsciously" (Libet, 1999, p. 47). This has confounded some philosophers; yet it is consistent with the picture we, Cisek (2007), and others have drawn: "The brain ... specifies ... potential actions that are currently available." That specification is measured as the RPs. As we have noted, consciousness is unnecessary and often counterproductive for these lower-level motor plans.

To press the button—or to do nothing—are offered as motor plans to the next level up, where consciousness decides; and can choose—yes or no (Libet, 1999).

Dispassionate contemplation of more portentous possibilities requires peace of mind. “Reason grants a hearing to both sides, then seeks to postpone action, even its own, in order that it may gain time to sift out the truth; but anger is precipitate” (Plato, I.xi, cited in Elster, 2004, p. 33); anger ain’t got no time to waste speculating. This is true of other strong emotions and dispositions, as well as of some traits, such as ADHD (Killeen, 2015; Killeen et al., 2012). The sense of urgency lowers the threshold for action (the vigilance parameter) so that practically the first perception/action match that comes up is acted upon. Sometimes this saves lives; sometimes it ends them in accidents and “crimes of passion.” In the latter cases, aggression is treated as a reflexive, feedforward action, giving the perpetrator no time for thought (*mens rea*), and partially absolving him of guilt (“It wasn’t a willful action, your honor; he just couldn’t help himself, and didn’t have time to think”). Had he given himself time for “a hearing for both sides” of action plans, the resulting instrumental act becomes purposeful, premeditated, willful, and fully punishable.

### ***Seeing that we are seeing***

There is a difference between seeing and seeing that you are seeing. The former needs no oversight; the latter constitutes conscious oversight, typically invoked when there is a mismatch signal. The mismatch beckons attention to recheck the representation that we first assigned our input, in yet greater detail—the vigilance parameter is raised—and to evaluate our action plans in light of that review. “Is that a leaf I am about to step on, or a little frog?” It is a review of candidate percepts and actions, in parallel, with the strongest candidates laid on the table of consciousness, either for action, or rejection for alternate candidates. Consciousness is the color that these actionable visions and plans are painted when they come into the focus of attention.

Sensation is tethered to the input; when that tether breaks, we hallucinate or misbehave (Jaynes, 1986; Killeen & Nash, 2003; Norman, 1981). When it succeeds, we have a map to the world, where we can approach our goal, or find new routes to it; or, failing that, select an alternate goal (“A *purposeful system* ... selects ends as well as means and thus displays *will*,” Ackoff, 1971, p. 666, ¶21). Consciousness need not always attend the details of such actions, where it might often “get in the way.” Jaynes (1986) calls selecting the goals for actions “struction,” without micromanaging their realization, as instructions would. When we are speaking, we may



formulate a communication goal for each sentence but do not consciously retrieve each word from memory and sequence it in a coherent fashion. Consciousness is neither needed nor wanted for that process—unless perchance we “find ourselves at a loss for words”: a mismatch has happened, and we pause to call out the big guns. “Like a bird’s life, [consciousness] seems to be made of an alternation of flights and perchings. The rhythm of language expresses this ...” (James, 1884, p. 2).

What is consciousness for? It is not for habitual behavior, the Type 1 tying of shoes. It arises when the shoestring breaks. At that point, we need options: Knot the laces? Find new ones? Strip laces from other shoes? Oh, of course, just use the other shoes for today. We picture these possibilities as though they were percepts, action plans, so to evaluate their utility. Consciousness is the stage on which these small dramas play out; the same stage that actions in the world play out until they become habitual. They often involve qualia—the untranslatable-into-words perceptions, desires, efforts, and feelings, that comprise an important part of our evaluations.

Why is this theater conscious? Why do we experience? The best answer may be “Why Not”? Why are we surprised that sentience may involve consciousness? Perhaps the “hard problem” of consciousness is awe gone unchecked: We confuse the ineffable, such as experience of the color red, with a special status (*qualia*), rather than with the shortcomings of our language in describing it. Language acquisition requires joint attention to objects (Tomasello & Farrar, 1986), available only by analogy for private stimuli (“red like an apple”). We may be in a situation similar to the analysis of subatomic phenomena, where symbols, whether language or mathematics, “do not follow the same rules as experience. They follow rules of their own. The problem is not *in* the language, the problem *is* the language.” How then do you communicate the experience? “You don’t. But by telling how you make quanta and how you measure them, you enable others to have it” (the physicist David Finkelstein, cited in Zukav, 2012, pp. 290–291). Like, “See this red apple.”

“Qualia are an aspect of subjectivity, of what goes on in a subject, and subjectivity is not epiphenomenal. On Earth, it is about the least epiphenomenal thing there is” (Godfrey-Smith, 2016). We manage our engagement with the world on the level of percepts, after all, not of sensations; of motor plans, not their realization. Often this requires not only seeing, but seeing that we are seeing, inspecting the thing seen, questioning it, and relating it to action. This plays out in the “controlled hallucinations” of consciousness.

Without conscious arbitration, there would be no way of knowing whether the percept on the table of interpretation came from the world, or from our history with it; whether that action plan might achieve a goal, or



if it is already being executed. Systems making simulations (Barsalou, 2009) that use the same networks as perceptions and actions, need a marker to help discriminate between imagined and realized. Conscious thought involves seeing that we are seeing, seeing that we are thinking, seeing that we are simulating. When the hard work of simulating and template matching is done, contemplation ceases, as attention turns to other occupations. We act. “There is no unbridgeable gap [between mind and matter: the ‘hard problem’ of qualia;] it is a problem that has to be dissolved, not solved. We hope we have contributed toward its dissolution” (Ginsburg & Jablonka, 2019, p. 482; who also speak for us).

Nobel laureate Roger Sperry (1966) shares similar thoughts, ones that echo Baer’s dialog of guests; Powers’s (1973) hierarchy of control systems; Miller et al.’s (1960) hierarchy of TOTE units; and Hebb’s (2018) hierarchy of cell-assemblies:

In my own hypothetical brain model, conscious awareness does get representation as a very real causal agent and rates an important place in the causal sequence and chain of control in brain events, in which it appears as an active, operational force. ... To put it very simply, it comes down to the issue of who pushes whom around in the population of causal forces that occupy the cranium. It is a matter, in other words, of straightening out the peck - order hierarchy among intracranial control agents. There exists within the cranium a whole world of diverse causal forces; what is more, there are forces within forces within forces, as in no other cubic half-foot of universe that we know. ... If one keeps climbing upward in the chain of command within the brain, one finds at the very top those over-all organizational forces and dynamic properties of the large patterns of cerebral excitation that are correlated with mental states .... Near the apex of this command system in the brain .... we find ideas.

We make ideas. And while there is a hierarchy, when we peek into the top floor of this command structure, we often find a colloquium of ideas, with quorum sensing to elevate the best in consciousness to action. A boardroom, not a monarch.

## Will

*Purposeful systems* can select ends as well as means, thereby displaying *will* (Ackoff, 1971, ¶21). That is a good definition, but something shy of a mechanism. Just how does such a system exert will? Can will coexist with determinism, or does it require indeterminism? Again, James’s thoughts enlighten the seeming dichotomy:

Indeterminism, on the contrary, says that the parts have a certain amount of loose play on one another, so that the laying down of one of them does not necessarily determine what the others shall be. It admits that possibilities may be in excess of actualities, and that things not yet revealed to our knowledge may really in themselves be ambiguous. Of two alternative futures which we conceive, both may

now be really possible; and the one becomes impossible only at the very moment when the other excludes it by becoming real itself. Indeterminism thus denies the world to be one unbending unit of fact. It says there is a certain ultimate pluralism in it; and, so saying, it corroborates our ordinary unsophisticated view of things. To that view, actualities seem to float in a wider sea of possibilities from out of which they are chosen; and, *somewhere*, indeterminism says, such possibilities exist, and form a part of truth. (James, 1896, p. 4)

Possibilities *are* in excess of actualities, one precluding the other “by becoming real itself.” The possibilities are voiced by Baer’s (1976) “response guests,” by Ainslie’s (1992) “interests,” by Grossberg’s (2021c) gated dipoles, by Sperry’s “causal forces.” By our maps of the world and our simulations of goals (Grush, 2004).

### **Determinism**

Even deterministic systems may “have a certain amount of loose play on one another.” Unpredictability—“things not yet revealed to our knowledge”—is not the same as indeterminacy. Consider the “logistic map” (May, 1976; Wikipedia, 2024), with the first value  $0 < x_1 < 1$ , and the next value of  $x$  chosen as:

$$x_{n+1} = rx_n(1 - x_n).$$

Iterate the equation, filling the computed value on the left into the right-hand side; and again; and again. What happens depends crucially on the value of  $r$ , ranging between 1 and 4, and for some of those values, the starting value of  $x$ ,  $x_1$ . For  $r$  between 1 and 3, the asymptotic value of  $x$  tends to  $(r-1)/r$  independent of  $x_1$ . For  $r$  above 3.57, however, the result is chaotic. The resulting values of  $x_n$  depend exquisitely on the values of  $r$  and  $x_1$ . For starting values of  $x_1$  differing in only the tenth decimal place, the trajectories will diverge to unpredictably different values. The only way to know where they land is to crank the iterative wheel.<sup>7</sup> “Unpredictability is not randomness, but in some circumstances looks very much like it” (Wikipedia, 2024). Determinism is not randomness, but in some circumstances looks very much like it.

Infinitely small differences in the starting weights on alternative action plans, weights that may vary with momentary vicissitudes in motivation, interest, or clouds before the sun, can affect what we ultimately choose to do. That neither we nor anyone else can predict what we will do, does not make it indeterminate; but for all practical purposes, it might as well. That we cannot predict what we might do does not make it an act of will; but for all practical purposes, when we decide to do it, it might as well.

The great Victorian physicist James Clerk Maxwell recognized these facts long ago: Initiating causes do not always precisely determine their effects:

When an infinitely small variation in the present state may bring about a finite difference in the state of the system in a finite time, the condition of the system is said to be unstable. It is manifest that the existence of unstable conditions renders impossible the prediction of future events, if our knowledge of the present state is only approximate, and not accurate. ... Stability is the characteristic of those systems from the contemplation of which determinists draw their arguments ...

[But identical] antecedents never again concur, and nothing ever happens twice. [It is said that:] ‘from like antecedents follow like consequents.’ But here we have passed from sameness to likeness, from absolute accuracy to a more or less rough approximation. ... At [some of] these points, influences whose physical magnitude is too small to be taken account of by a finite being, may produce results of the greatest importance. (Maxwell, 1882, pp. 442–443)

There is an interesting antiparallel between classical physical laws and deterministic proclamations. We know that the basic laws of mechanics are time-reversible—substitute  $-t$  for  $t$  everywhere in those equations and they remain valid. A video of a person tipping over a tall stack of bricks is unremarkable. Played in reverse, however, it is incredible. Both may be described by the same laws of mechanics. Why the difference in psychological impact? Physical laws need initial conditions—here mass, position, velocity, coefficients of friction, etc. The forward video is consistent with a simple specification: velocity zero, and a wide allowable range for the other variables. The reverse video requires an exquisitely precise determination of initial position and momenta.<sup>8</sup> Much more information is required to instantiate the reverse trajectories than the forward ones. They are subject to Maxwell’s “influences whose physical magnitude is too small to be taken account of by a finite being.”

Just as the laws of physics are time-asymmetric, working better in forward than reverse, the laws of determinism are time-asymmetric, working better in reverse than in forward. Having acted, we can be sure causes occurred, and often find the salient ones. But, looking forward, given potential causes, the ensuing action is often unpredictable. “At these [inflection] points, influences whose physical magnitude is too small to be taken account of by a finite being, may produce results of the greatest importance.” Like the shadow of a smile.

Epistemic unpredictability does not, of course, constitute evidence for ontological indeterminism, much less for “agency-between-the-cracks.” It does suggest, however, that despite the manifold success of the assumption of determinism in the sciences, in complex systems—whether turbulent flow or turbulent thoughts—the claim of determinism is typically cashed out only in retrospect. It is an act of faith going forward, a generalization from its great effectiveness in analyses of “stable systems,” to complex systems; justifiable statistically at best, as a “probabilistic determinism” (Glynn, 2010; Strevens, 2011). In the end, determinism does not entail

determinability: “the biosphere does not contain a predictable class of objects or of events but constitutes a particular occurrence, compatible indeed with first principles, but not deducible from those principles and therefore essentially unpredictable” (Monod, 1974, p. 43).

Does that leave a place for free will anywhere amongst the swerving atoms?

### Free will?

What we call agency has for millennia been called the *will*. An agent’s action counts as free unless coerced, caused by another agent’s use or threat of force.<sup>9</sup> That is the common understanding of freedom: It is ambit for self-motivated and uncoerced behavior. “Freedom in all these senses presents simply no problem at all. No matter what the soft determinist means by it,—whether he means the acting without external constraint [etc.] ...—who cannot answer him that sometimes we are free and sometimes we are not?” (James, 1896, p. 2). Its opposite is behavior motivated by threat of punishment, limitations on what you *may* do: What other agents will *permit* you to do in your efforts to attain your goals. Moral proscripts and civic laws punish deviation, limiting what you may do (with impunity). Beyond these often-wide ambits, slavery and other forms of subjection consist of what others *require* you to do, whether you want to or not, to avoid punishment. The top controllers in these cases are outside your body, even though you will have internalized some of them, especially those formed at your parents’ knees, or under their straps, in inhibitions that Freud called the superego.

No agent can be free from, or independent of, her own desires and emotions; nor is that necessary for freedom, nor is it desirable. Simone de Beauvoir taught this to her partner, Jean Paul Sartre: “Being free does not mean that you must cease to wish for things.” Indeed, being free means being allowed to pursue the things that you wish for. Freedom, in the end, is political, not metaphysical or psychological. It can’t be found in the soul, or in the brain; but only in interactions with other people, and with the polity (Gazzaniga, 2012a, 2012b).

*Free will* then is a misnomer for agency, ability to choose our next move. All of the moves we have made up to this moment have been caused by the forces just at play, whatever they were, interacting with our system. Our behavior was caused, even if the regnant causal factor is undeterminable. All of our next moves will have been determined by some of the forces at play when we made them. These include how you thought about them, how you simulated them, your bodily dispositions at the time, your recent history, and on occasion the passing of a cloud before the sun, or

your flipping of a coin. Taking a path less traveled, and its unexpected roadside attractions, these can make all the difference. Such things cannot be predicted, or even iteratively simulated. You do what you want to do, at that moment, and you can change what you want to do, if you want to, by placing the very light weight of attention on the scales of the available options. Which stack captures your attention is often as fickle as the flip of a coin, as inscrutable as Maxwell's "influences whose physical magnitude is too small to be taken account of by a finite being"; yet the direction of your attention is a core aspect of your agency, as it breeds involvement. It amplifies the power of minuscule propensities, in the end toppling one stack of hierarchies in favor of another. Your wants are dependent on causal factors, both internal and external; these can be as complex as the reading of this paper, or as simple as someone saying, "Please, don't go." Think about it. Your thoughts themselves will roll the deterministic dice. "A free choice is one that has been made for one's own reasons" (Shepard, 2008, p. 27); inscrutable though they may be. This is the only kind of free will worth wanting (Dennett, 2015).

## **Responsibility**

How can we hold people responsible if their behavior is determined? No problem, if that is what our society chooses to do. Punishment cannot undo the past, of course, "A sensible person will not punish a person because he has sinned, but in order to keep him from sin; for while the past may not be recalled, the future may be forestalled" (Plato, I.xix, cited in Elster, 2004, p. 32). As retribution, punishment can also hurt the miscreant, which may please those in control or those offended by his behavior.

As Dennett (2017, p. 227) notes, "our zealous search for 'justice' is often little more than our instinctual yearning for retaliation dressed up to look respectable." Sapolsky (2023) spends the last third of his brilliant book discussing how to live a humane life in a deterministic universe—ideally, one without retaliation. But punishment *can* be justified in terms of its consequences: "Punishment can be fair, punishment can be justified, and in fact, our societies could not manage without it" (Dennett, 2017, p. 228). The yellow card in soccer discourages unsportsmanlike behavior both by that player and by his teammates. Speeding tickets save lives. The past is determined, the future is open; the future can be changed, changed by rewards, by punishment, and by information. There is no predicting how they will come at you, or the effect that they will have on you. "It is literally *a world of possibilities*" (Dickinson, 2005; Hocutt, 2019; Mitchell, 2023, p. 131).

## The mind's I

Who am I, then, who seems to be choosing what to look at, what to think, what to do? Who is it that seems to be willing; or not? There isn't a singular I. Its semblance emerges from the intermittent and momentary intervention on the highest levels of the control systems, as described by Varela et al. (1993):

The apparent totality and continuity of consciousness masks the discontinuity of momentary consciousness related to one another by cause and effect. A traditional metaphor for this illusory continuity is the lighting of one candle with a second candle, a third candle from that one, and so on—the flame is passed from one candle to the next without any material basis being passed on (p. 69).

The brain makes models to make sense of the world. When considering the play of perception, attention, and proto action that constitutes thought, what learned category will come closest to resonating with that constellation of events? That must be the brain's model of me at large. I walk around the yard on a sunny morning to admire the spring flowers, then I decide to turn back to press a second cup of coffee and glance at the morning news. This is Peter behaving. It is a good model of him, resonant in many ways with Peter thinking. It resonates because I will often think such things, then do such things. My brain *must* have a model of me, a causal model, as knowing which way I am looking at a street crossing makes the difference between a safe and fatal next step (Jékely et al., 2021). The brain uses Peter behaving as the avatar for Peter thinking. The little man inside is a caricature of the big man it is inside. Peter's I is an aggregate of the experiences of the control system, tagged for retrieval as his Self.

## Summary

There are several radically different views of a behaving organism: As a rudderless raft in the stormy seas of nature, a locus at which forces converge and then diverge, a lens through which energy passes, a marionette moved by historic strings, unseen. Versions are found in Skinner, and in interbehaviorism (Hayes & Fredericks, 1999). Elements are found in radical monism (Hayes & Fryling, 2014), in molar behaviorism (Baum, 2011), in teleological behaviorism (Rachlin, 2014), in contextual behavior science (Hayes, 1993), in RFT (Hayes et al., 1993; Stewart, 2017), and in Sapolsky's (2023) erudite and forceful defense of determinism.

The contrasting view is that we are purposeful and goal-driven organisms—agents. That we think to act more effectively (Shepard, 2008). Versions abound in the larger field of psychology and humanities, brought to an evolutionary head in Mitchell's (2023) *Free Agents*. It is, surprisingly, also found in the literature of behaviorists: in Skinner (Zuriff, 1985), in

Baer (1976), in contextual behavioral science (Hayes et al., 2012; Hayes & Wilson, 2003; Zettle et al., 2016), in radical behaviorism (Chiesa, 1994; Moore, 1981, 2008), in emergent behaviorism (Killeen, 1984), in picoeconomics (Ainslie, 1992, 2001) in purposive behaviorism (Tolman, 1932), and in intentional behaviorism (Foxall, 2007, 2008; Marr, 2007). Kantor's holistic interbehaviorism, involving "mutual reciprocal and simultaneous interactions between response functions and stimulus functions" (Morris, 1982, p. 191), constitutes a complex system defying analysis in terms of deterministic causal chains.<sup>10</sup> The interacting "actors" in these plays have different names—covert stimulus-response-sequences, guests, interests, private behaviors, thoughts, control systems, TOTE units, cell-assemblies—but they are largely reading the same script, and making similar moves. With this pandemonium at play, it is often impossible to predict which interest will be heard the loudest; all are possible. Some voices capture attention, and by being attended, their voice grows more salient. The conferees call for closure to the deliberation, whether from conviction or merely satiation/fatigue in ideation of the options: they satisfice. The last one standing wins; we act.

The first vision is called "incompatibilist determinist"—determinism with no room left for autonomous agency, exemplified by Sapolsky (2023). The second is "incompatibilist libertarian," exemplified by Mitchell (2023). A third is compatibilist, allowing both determinism and agency, exemplified by Dennett (2015). Our position is a temporally asymmetric kind of compatibilism: Libertarian looking forward, deterministic looking back. We reject the Laplacian hypothesis that knowing all positions and momenta of all particles we could predict all possible futures. We cannot know those givens, we cannot reduce the uncertainty of their joint positions and momenta below Planck's constant. (And we wonder who would be "given" this information, what the byte size is on Her computer below which She must truncate, and whether She herself is part of the system analyzed.)

Sapolsky nonetheless makes a convincing case for all of the ways our biology and history as individuals influence our choices. These influences, these causal factors, are sometimes so strong as to be determinative. But not always; there is often room for some Planck slippage. Whether the random factor is in our environment or in our electrons, the future is only probabilistically predictable. Play "rock, scissors, paper" with a friend for a homely example. The thesis of determinism of future events is neither provable nor disprovable; and therefore it is not scientific. Whereof one cannot speak, thereof one must not pontificate. We reject the hypothesis.

Concerning *fait accompli*, the failure to find their determinants, whichever of many they may have been, is not an argument against their existence. As scientists, we favor the act of faith that all realized events were



determined by prior conditions; whatever, and however inscrutable, they might have been. This deterministic asymmetry is not exceptional: It is akin to the breakdown of the quantum wavefunction upon observation, with probabilities before the act, resolved into a particle moving through space and time after it. This framing serves the cake of our mechanistic science looking back while letting us have its frosting of personal agency looking ahead.

We admire Skinner's assertion that the operant is the essence of purposeful behavior but prefer Thorndike's version of the law of effect. Thorndike's treatment of the operant as goal-directed behavior is best represented as a control system, the bottom diagram in [Figure 1](#). But that simple diagram supports only simple behavior. That is only a partial model; additional levels of controllers modulate the controllers above and below. The more complex behavior necessary in a complex environment requires that we move to the bottom of [Figure 2](#). There is precedent for hierarchical organization in biology (Dawkins, 1976), in the brain (Frith, 2009; Négyessy et al., 2006), and in complex systems in general (Simon, 1962). No fear that this conception must lead to an infinite regress, for even infinite series can converge to finite limits: The sum of  $2^{-n}$  from  $n = 0$  to infinity is a humble 2. According to the ratio test, a series converges when the  $n$ th + 1 term is less than the  $n$ th term. Here, this requires that the information processed by a higher level is less than that processed by the one below. The trigger "Go" carries less information than that required to actually move the arm. The subordinate systems process that information. It works. The architectures of ART incorporate such control systems while providing a richer and more biologically motivated system of systems.

Who is the controller at the top of the stack? We call it I. I deliberate among my response guests, evaluating their simulations—telling them to "put up [with an attractive percept or action plan] or shut up." I is not a ghost in the machine, but rather the percept that my brain assigns to its model of me behaving, both overtly and covertly. My avatar. Consciousness abets this process by laying out the playing field as percepts and actions and evaluating their hedonic prospects, with the process itself represented as I thinking. Self-knowledge is sometimes counter-productive to those ends (Greenwald, 1997; Krebs et al., 1988), and oftentimes irrelevant to it; the job of the I is to attain ever changing desiderata, not to mull them over when not required, nor to chat about them. Positive illusions (Taylor et al., 2000), such as determinism and free will, may be essential to achieving the agent I's various desiderata. I think that is just fine, as I have evolved to behave for the proxies of evolutionary success, not for fact checkers. I can even turn a blind I, when I wish, to Frith's observation that "The top of top-down control is not to be found in the individual brain, but in the



culture that is the human brain's unique environmental niche.” (Frith, 2009, p. 199).

Agency, then, is not one thing, but an orchestra of things. It is a history of interactions in the world, imprinting possible percepts and action plans—the voices in the choir. They sometimes sing *sub rosa*, guiding habitual patterns without calling attention. But when surprising themes evolve from their harmonies, dissonances unscored, the cacophony over-tops our vigilance threshold to capture attention. The music rises to the playing field of consciousness and, time allowing, simulations are run—to jump or to swerve? To the mountains or to the shore?—with the hedonic implications of these deliberations sampled. Those that are surveyed the longest, because most enticing, tend to win. Attention dims, and the last standing deliberation becomes a new desideratum; lower-level action plans are engaged. I swerve to the beach. This choice is agency.

## Acknowledgments

We thank Art Glenberg, Mary Titus, two erudite and helpful anonymous reviewers for comments on the Ms. We thank David Trafimow for nurturing this project.

## Author notes

Killeen received his Doctorate under BF Skinner at Harvard. He is now Professor of Psychology Emeritus at ASU.

Helms Tillery received his Doctorate under Soechting and Ebner at the University of Minnesota. He is now Lincoln Professor of Neural Engineering Research and Ethics at ASU.

Cabrera received his Doctorate from UdG. He currently works as a Senior Researcher at the Center for Research in Comparative Behavior and Cognition at UdG.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Notes

1. These are templates for common potential perceptions in that context. They may be hard-wired, as in perceptual illusions, or learned, as is the template for “doctor” in the context of “hospital” and “nurse.” Like a Bayesian prior, they can be updated, but they do not generally have the statistical properties of Bayesian priors. They are the ideas served up by generative AI. They are an instance of Ginsburg and Jablonka (2019) “categorizing sensory states”.
2. The definitions in the following paragraphs come directly from this article.
3. Here, “open” means that the set points or inputs to the system are modifiable. This is not to be confused with open-loop systems, where the success of the performance has no effect on the system.

4. Many of the observations in this section issue from discussions with Max Hocutt.
5. Within bounds. Resolution of perceptions, or satisfaction of appetites, will change the sign of the feedback, from positive to zero or even negative; the fresh salad is attractive, but, once consumed, a second or third would be repulsive.
6. There are many points of convergence between Mitchell's book and this paper.
7. And even then, prediction is limited by the finite precision of the number of bits of representation available in your computer. See Hoover and Hoover (2012).
8. And, of course, there is the mysterious force that animates them to fly into a column, each with their own accurate momentum, aggregating there like a purposive slime mold. Example from Sachs (1987). Falling, the stack of bricks exemplify the simple reaction in the top of Figure 2; Reassembling, the purposive machines in the bottom of that figure; abilities we do not credit to bricks, especially when bumped by other bricks enroute to their "goal". Such improbability is one of the causes of the awe that many experienced watching the Space X booster return to land upright on its tripod.
9. These ideas issue from discussions with Max Hocutt. For more extensive analysis, see his *Grounded Ethics* (2000).
10. Araiba (2020) attributes this plethora of approaches to the problem of agency, as due to adaptive radiation of behavioral theory to niches occupied by cognitive and clinical psychologists.

## Funding

The author(s) reported there is no funding associated with the work featured in this article.

## ORCID

Peter R. Killeen  <http://orcid.org/0000-0002-0889-4040>

Stephen Helms Tillery  <http://orcid.org/0000-0001-9938-8655>

Felipe Cabrera  <http://orcid.org/0000-0002-6015-1957>

## References

- Ackoff, R. L. (1971). Towards a system of systems concepts. *Management Science*, 17(11), 661–671. <https://doi.org/10.1287/mnsc.17.11.661>
- Adami, C. C. (2007). Who watches the watcher? *Science*, 316(5828), 1125–1126. <https://doi.org/10.1126/science.1141809>
- Ainslie, G. (1986). Beyond microeconomics: Conflict among interests in a multiple self as a determinant of value. In J. Elster (Ed.), *The multiple self* (pp. 133–175). Cambridge University Press.
- Ainslie, G. (1992). *Picoeconomics*. Cambridge University Press.
- Ainslie, G. (2001). *Breakdown of will*. Cambridge University Press.
- Ainslie, G. (2005). Precis of breakdown of will. *The Behavioral and Brain Sciences*, 28(5), 635–650; discussion 650–673. <https://doi.org/10.1017/S0140525X05000117>
- Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial Intelligence*, 149(1), 91–130. [https://doi.org/10.1016/S0004-3702\(03\)00054-7](https://doi.org/10.1016/S0004-3702(03)00054-7)
- Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain*. MIT Press.
- Anscombe, G. E. M. (2011). *Human life, action and ethics: Essays by GEM Anscombe* (Vol. 4). Andrews UK Limited.

- Ashby, F. G., Zetzer, H. A., Conoley, C. W., & Pickering, A. (2024). Just do it: A neuro-psychological theory of agency, cognition, mood, and dopamine. *Journal of Experimental Psychology. General*, 153(6), 1582–1604. <https://doi.org/10.1037/xge0001587>
- Åström, K. J., & Murray, R. M. (2010). *Feedback systems: An introduction for scientists and engineers*. Princeton University Press.
- Baer, D. M. (1976). The organism as host. *Human Development*, 19(2), 87–98. <https://doi.org/10.1159/000271519>
- Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 364(1521), 1281–1289. <https://doi.org/10.1098/rstb.2008.0319>
- Bartlett, F. C. (1995). *Remembering: A study in experimental and social psychology*. Cambridge University Press.
- Baum, W. M. (1973). The correlation-based law of effect. *Journal of the Experimental Analysis of Behavior*, 20(1), 137–153. <https://doi.org/10.1901/jeab.1973.20-137>
- Baum, W. M. (2011). Behaviorism, private events, and the molar view of behavior. *The Behavior Analyst*, 34(2), 185–200. <https://doi.org/10.1007/BF03392249>
- Baum, W. M. (2018). Three laws of behavior: Allocation, induction, and covariance. *Behavior Analysis: Research and Practice*, 18(3), 239–251. <https://doi.org/10.1037/bar0000104>
- Bechtel, W. (2013). The endogenously active brain: The need for an alternative cognitive architecture. *Philosophia Scientiae*, 17(2), 3–30. <https://doi.org/10.4000/philosophiascientiae.846>
- Bechtel, W. (2019). Resituating cognitive mechanisms within heterarchical networks controlling physiology and behavior. *Theory & Psychology*, 29(5), 620–639. <https://doi.org/10.1177/0959354319873725>
- Beer, R. D. (2004). Autopoiesis and cognition in the game of life. *Artificial Life*, 10(3), 309–326. <https://doi.org/10.1162/1064546041255539>
- Behnke, E. A. (1997). Ghost gestures: Phenomenological investigations of bodily micro-movements and their intercorporeal implications. *Human Studies*, 20(2), 181–201. <https://doi.org/10.1023/A:1005372501258>
- Bogdan, R. (2013). *Roderick M. Chisholm* (Vol. 7). Springer Science & Business Media.
- Borgstede, M., & Eggert, F. (2021). The formal foundation of an evolutionary theory of reinforcement. *Behavioural Processes*, 186, 104370. <https://doi.org/10.1016/j.beproc.2021.104370>
- Bourbon, W. T. (1996). On the accuracy and reliability of predictions by perceptual control theory: Five years later. *The Psychological Record*, 46(1), 39–47. <https://doi.org/10.1007/BF03395162>
- Braitenberg, V. (1986). *Vehicles: Experiments in synthetic psychology*. MIT Press.
- Breland, K., & Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, 16(11), 681–684. <https://doi.org/10.1037/h0040090>
- Burgos, J. E., & Killeen, P. R. (2019). Suing for peace in the war against mentalism. *Perspectives on Behavior Science*, 42(2), 241–266. <https://doi.org/10.1007/s40614-018-0169-2>
- Canfield, J. V. (Ed.). (1966). *Purpose in nature*. Prentice Hall.
- Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., Chialvo, D. R., & Nutt, D. (2014). The entropic brain: A theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in Human Neuroscience*, 8, 20. <https://doi.org/10.3389/fnhum.2014.00020>
- Catania, A. C. (2013). *Learning* (5th ed.). Sloan Publishing.

- Chiesa, M. (1994). *Radical behaviorism: The philosophy and the science*. Authors Cooperative.
- Cisek, P. (2007, September 29). Cortical mechanisms of action selection: the affordance competition hypothesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362(1485), 1585–1599. <https://doi.org/10.1098/rstb.2007.2054>
- Cowie, S., & Davison, M. (2016, March). Control by reinforcers across time and space: A review of recent choice research. *Journal of the Experimental Analysis of Behavior*, 105(2), 246–269. <https://doi.org/10.1002/jeab.200>
- Davidson, D. (2001). *Essays on actions and events: Philosophical essays* (Vol. 1). Oxford University Press on Demand.
- Davison, M. (2017). Killeen and Jacobs are not wrong. *The Behavior Analyst*, 40(1), 57–64. <https://doi.org/10.1007/s40614-017-0118-5>
- Dawkins, R. (1976). Hierarchical organisation: A candidate principle for ethology. In P. P. G. Bateson & R. A. Hinde (Eds.), *Growing points in ethology* (pp. 7–54). Cambridge University Press.
- De Chardin, P. T., Huxley, J., & Wall, B. (1965). *The phenomenon of man*. Collins London.
- Dennett, D. C. (1975). Why the law of effect will not go away. *Journal for the Theory of Social Behaviour*, 5(2), 169–188. <https://doi.org/10.1111/j.1468-5914.1975.tb00350.x>
- Dennett, D. C. (1995). Darwin's dangerous idea. *The Sciences*, 35(3), 34–40. <https://doi.org/10.1002/j.2326-1951.1995.tb03633.x>
- Dennett, D. C. (2015). *Elbow room, new edition: The varieties of free will worth wanting*. MIT Press.
- Dennett, D. C. (2017). Reflections on Sam Harris' "Free Will". *Rivista Internazionale di Filosofia e Psicologia*, 8(3), 214–230.
- Dennett, D., & Akins, K. (2008). Multiple drafts model. *Scholarpedia*, 3(4), 4321. <https://doi.org/10.4249/scholarpedia.4321>
- Dickinson, E. (2005). *The poems of Emily Dickinson: Reading edition (poem 466)*. Belknap Press.
- Domjan, M. (2005). Pavlovian conditioning: A functional perspective. *Annual Review of Psychology*, 56(1), 179–206. <https://doi.org/10.1146/annurev.psych.55.090902.141409>
- Domjan, M., Cusato, B., & Villarreal, R. (2000). Pavlovian feed-forward mechanisms in the control of social behavior. *The Behavioral and Brain Sciences*, 23(2), 235–249; discussion 249–282. <https://doi.org/10.1017/s0140525x00002430>
- Donahoe, J. W. (2014). Evocation of behavioral change by the reinforcer is the critical event in both the classical and operant procedures. *International Journal of Comparative Psychology*, 27(4), 537–543. <https://doi.org/10.46867/ijcp.2014.27.04.05>
- Donahoe, J. W., & Vegas, R. (2004, January). Pavlovian conditioning: The CS-UR relation. *Journal of Experimental Psychology. Animal Behavior Processes*, 30(1), 17–33. <https://doi.org/10.1037/0097-7403.30.1.17>
- Dworkin, B. R. (1993). *Learning and physiological regulation*. University of Chicago Press.
- Ellis, G. F. R. (2012). Top-down causation and emergence: Some comments on mechanisms. *Interface Focus*, 2(1), 126–140. <https://doi.org/10.1098/rsfs.2011.0062>
- Elster, J. (2004). Emotions and rationality. From feelings and emotions. In A. S. R. Mansteadt, N. Frijda, & A. Fischer (Eds.), *The Amsterdam symposium* (pp. 30–57). Cambridge University Press.
- Fanselow, M. S. (1997). Species-specific defense reactions: Retrospect and prospect. In M. E. Bouton & M. S. Fanselow (Eds.), *Learning, motivation, and cognition: The functional behaviorism of Robert C. Bolles* (pp. 321–341). American Psychological Association. <https://doi.org/10.1037/10223-016>

- Foxall, G. R. (2007). Intentional behaviorism. *Behavior and Philosophy*, 35, 1–55. [https://doi.org/10.1007/978-3-030-77395-3\\_13](https://doi.org/10.1007/978-3-030-77395-3_13)
- Foxall, G. R. (2008). Intentional behaviorism revisited. *Behavior and Philosophy*, 36, 113–155.
- Frith, C. D. (2009). Free will top-down control in the brain. In N. Murphy, G. F. R. Ellis, & T. O'Connor (Eds.), *Downward causation and the neurobiology of free will* (pp. 199–209). Springer.
- Gardner, R. A., & Gardner, B. T. (1988). Feedforward versus feedbackward: An ethological alternative to the law of effect. *Behavioral and Brain Sciences*, 11(3), 429–447. <https://doi.org/10.1017/S0140525X00058258>
- Garrigan, P., & Kellman, P. J. (2008). Perceptual learning depends on perceptual constancy. *Proceedings of the National Academy of Sciences of the United States of America*, 105(6), 2248–2253. <https://doi.org/10.1073/pnas.0711878105>
- Gazzaniga, M. (2012a). Free Will is an illusion, but you're still responsible for your actions. *The Chronicle of Higher Education*, 1–6. Retrieved from <https://www.chronicle.com/article/free-will-is-an-illusion-but-youre-still-responsible-for-your-actions/>.
- Gazzaniga, M. (2012b). *Who's in charge?: Free will and the science of the brain*. Hachette UK.
- Ghirlanda, S. (2018). ecco: An error correcting comparator theory. *Behavioural Processes*, 154, 36–44. <https://doi.org/10.1016/j.beproc.2018.03.009>
- Gibson, J. J. (1962). Observations on active touch. *Psychological Review*, 69(6), 477–491. <https://doi.org/10.1037/h0046962>
- Ginsburg, S., & Jablonka, E. (2019). *The evolution of the sensitive soul: Learning and the origins of consciousness*. MIT Press.
- Glynn, L. (2010). Deterministic chance. *The British Journal for the Philosophy of Science*, 61(1), 51–80. <https://doi.org/10.1093/bjps/axp020>
- Godfrey-Smith, P. (2016). Individuality, subjectivity, and minimal cognition. *Biology & Philosophy*, 31(6), 775–796. <https://doi.org/10.1007/s10539-016-9543-1>
- Good, I. J. (1961). A causal calculus. *The British Journal for the Philosophy of Science*, 11(44), 305–318. <https://doi.org/10.1093/bjps/XI.44.305>
- Gould, S. J., & Vrba, E. S. (1982). Exaptation—A missing term in the science of form. *Paleobiology*, 8(1), 4–15. <https://doi.org/10.1017/S0094837300004310>
- Greenwald, A. G. (1997). Self-knowledge and self-deception: Further consideration. In M. S. Myslobodsky (Ed.), *The mythomanias: The nature of deception and self-deception* (pp. 51–72). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Grossberg, S. (1980). How does the brain build a cognitive code. *Psychological Review*, 87(1), 1–51. [https://doi.org/10.1007/978-94-009-7758-7\\_1](https://doi.org/10.1007/978-94-009-7758-7_1)
- Grossberg, S. (2012). *Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control* (Vol. 70). Springer Science & Business Media.
- Grossberg, S. (2021a). *Conscious mind, resonant brain: How each brain makes a mind*. Oxford University Press.
- Grossberg, S. (2021b). Toward autonomous adaptive intelligence: Building upon neural models of how brains make minds. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(1), 51–75. <https://doi.org/10.1109/TSMC.2020.3041476>
- Grossberg, S. (2021c). A unified neural theory of conscious seeing, hearing, feeling, and knowing. *Cognitive Neuroscience*, 12(2), 69–73. <https://doi.org/10.1080/17588928.2020.1839401>
- Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *The Behavioral and Brain Sciences*, 27(3), 377–396; discussion 396–442. <https://doi.org/10.1017/s0140525x04000093>

- Hanson, C., Caglar, L. R., & Hanson, S. J. (2018). Attentional bias in human category learning: The case of deep learning. *Frontiers in Psychology*, 9(374), 374. <https://doi.org/10.3389/fpsyg.2018.00374>
- Hanson, S. J. (1990). A stochastic version of the delta rule. *Physica D: Nonlinear Phenomena*, 42(1–3), 265–272. [https://doi.org/10.1016/0167-2789\(90\)90081-Y](https://doi.org/10.1016/0167-2789(90)90081-Y)
- Hanson, S. J., & Burr, D. J. (1990). What connectionist models learn: Learning and representation in connectionist networks. *Behavioral and Brain Sciences*, 13(3), 471–489. <https://doi.org/10.1017/S0140525X00079760>
- Hanson, S. J., & Negishi, M. (2002). On the emergence of rules in neural networks. *Neural Computation*, 14(9), 2245–2268. <https://doi.org/10.1162/089976602320264079>
- Hanson, S. J., & Timberlake, W. (1983). Regulation during challenge: A general model of learned performance under schedule constraint. *Psychological Review*, 90(3), 261–282. <https://doi.org/10.1037/0033-295X.90.3.261>
- Hayes, L. J., & Fredericks, D. W. (1999). Interbehaviorism and interbehavioral psychology. In W. O'Donohue & R. Kitchener (Eds.), *Handbook of behaviorism* (pp. 71–96). Academic.
- Hayes, L. J., & Fryling, M. J. (2014). Motivation in behavior analysis: A critique. *The Psychological Record*, 64(2), 339–347. <https://doi.org/10.1007/s40732-014-0025-z>
- Hayes, S. C. (1993). *Analytic goals and the varieties of scientific contextualism*. Context Press.
- Hayes, S. C., & Wilson, K. G. (2003). Mindfulness: Method and process. *Clinical Psychology: Science and Practice*, 10(2), 161–165. <https://doi.org/10.1093/clipsy.bpg018>
- Hayes, S. C., Barnes-Holmes, D., & Wilson, K. G. (2012). Contextual behavioral science: Creating a science more adequate to the challenge of the human condition. *Journal of Contextual Behavioral Science*, 1(1–2), 1–16. <https://doi.org/10.1016/j.jcbs.2012.09.004>
- Hayes, S. C., Hayes, L. J., Reese, H., W., & Sarbin, T. R. (Eds.). (1993). *Varieties of scientific contextualism*. Context Press.
- He, D., & Ogmen, H. (2021). Sensorimotor self-organization via circular-reactions. *Frontiers in Neurobotics*, 15, 658450. <https://doi.org/10.3389/fnbot.2021.658450>
- Hebb, D. O. (2018). Elaborations of Hebb's cell assembly theory. In *Neuropsychology after Lashley* (pp. 483–496). Routledge.
- Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, 56(5), 872–876. <https://doi.org/10.1037/h0040546>
- Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences*, 6(6), 242–247. [https://doi.org/10.1016/S1364-6613\(02\)01913-7](https://doi.org/10.1016/S1364-6613(02)01913-7)
- Hocutt, M. (1974). Aristotle's four because. *Philosophy*, 49(190), 385–399. <https://doi.org/10.1017/S0031819100063324>
- Hocutt, M. (2017). Just responsibility. *Behavior and Philosophy*, 45, 79–89. <https://www.jstor.org/stable/90018265>
- Hocutt, M. (2019). Behaviorist agency. *Behavior and Philosophy*, 47, 67–80. <https://www.jstor.org/stable/26921923>
- Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An eternal golden braid*. Basic Books.
- Hofstadter, D. R. (1982). Who shoves whom around inside the careenium? Or what is the meaning of the word “I”? *Synthese*, 53(2), 189–218. <https://doi.org/10.1007/BF00484897>
- Hofstadter, D. R. (2007). *I am a strange loop*. Basic Books.
- Holland, J. H. (1995). *Hidden order*. Addison-Wesley.
- Holland, P. C. (1979). The effects of qualitative and quantitative variation in the US on individual components of Pavlovian appetitive conditioned behavior in rats. *Animal Learning & Behavior*, 7(4), 424–432. <https://doi.org/10.3758/BF03209696>



- Holland, P. C. (1980). Influence of visual conditioned stimulus characteristics on the form of Pavlovian appetitive conditioned responding in rats. *Journal of Experimental Psychology. Animal Behavior Processes*, 6(1), 81–97.
- Hollis, K. L. (1997). Contemporary research on Pavlovian conditioning. A “new” functional analysis. *The American Psychologist*, 52(9), 956–965. <https://doi.org/10.1037/0003-066x.52.9.956>
- Honey, C. J., Mahabal, A., & Bellana, B. (2023). Psychological momentum. *Current Directions in Psychological Science*, 32(4), 284–292. <https://doi.org/10.1177/09637214221143053>
- Hoover, W. G., & Hoover, C. G. (2012). *Time reversibility, computer simulation, algorithms, chaos* (Vol. 13). World Scientific.
- Hull, D. L. (1974). *Philosophy of biological science*. Prentice-Hall.
- Hume, D. (1748). Of liberty and necessity. In D. Hume (Ed.), *An enquiry concerning human understanding*. <https://is.gd/dPAths>. Reprinted from <https://is.gd/dPAths>
- James, W. (1879). Are we automata? *Mind*, os-4(13), 1–22. <https://doi.org/10.1093/mind/os-4.13.1>
- James, W. (1884). On some omissions of introspective psychology. *Mind*, os-IX(33), 1–26. <https://doi.org/10.1093/mind/os-IX.33.1>
- James, W. (1896). The dilemma of determinism. In W. James (Ed.), *The will to believe and other essays in popular philosophy* (pp. 145–183). Longmans, Green and Co. <https://doi.org/10.1037/11061-005>
- Jaynes, J. (1986). Consciousness and the voices of the mind. *Canadian Psychology*, 27(2), 128–148. <https://doi.org/10.1037/h0080053>
- Jékely, G., Godfrey-Smith, P., & Keijzer, F. (2021, March 29). Reafference and the origin of the self in early nervous system evolution. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 376(1821), 20190764. <https://doi.org/10.1098/rstb.2019.0764>
- Juarrero, A. (2000). Dynamics in action: Intentional behavior as a complex system. *Emergence*, 2(2), 24–57. [https://doi.org/10.1207/S15327000EM0202\\_03](https://doi.org/10.1207/S15327000EM0202_03)
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Killeen, P. R. (1981). Learning as causal inference. In M. Commons & J. A. Nevin (Eds.), *Quantitative studies of behavior* (pp. 289–312). New York: Pergamon.
- Killeen, P. R. (1984). Emergent behaviorism. *Behaviorism*, 12(2), 25–39.
- Killeen, P. R. (2001, August). The four causes of behavior. *Current Directions in Psychological Science*, 10(4), 136–140. <https://doi.org/10.1111/1467-8721.00134>
- Killeen, P. R. (2014). Pavlov + Skinner = Premack. *International Journal of Comparative Psychology*, 27(4), 544–568. <https://doi.org/10.46867/ijcp.2014.27.04.04>
- Killeen, P. R. (2015). Models of ADHD: Five ways smaller sooner is better. *Journal of Neuroscience Methods*, 252, 2–13. <https://doi.org/10.1016/j.jneumeth.2015.01.011>
- Killeen, P. R. (2023). Theory of reinforcement schedules. *Journal of the Experimental Analysis of Behavior*, 120(3), 289–319. <https://doi.org/10.1002/jeab.880>
- Killeen, P. R., & Glenberg, A. M. (2010). Resituating cognition. *Comparative Cognition & Behavior Reviews*, 5, 59–77. <https://doi.org/10.3819/ccbr.2010.50003>
- Killeen, P. R., & Jacobs, K. W. (2017). The modulated contingency. *The Behavior Analyst*, 40, 1–20. <https://doi.org/10.1007/s40614-017-0101-1>
- Killeen, P. R., & Nash, M. (2003). The four causes of hypnosis. *The International Journal of Clinical and Experimental Hypnosis*, 51(3), 195–231. <https://doi.org/10.1076/iceh.51.3.195.15522>



- Killeen, P. R., Tannock, R., & Sagvolden, T. (2012). The four causes of ADHD: A framework. In S. C. Stanford & R. Tannock (Eds.), *Behavioral neuroscience of attention deficit hyperactivity disorder and its treatment* (Vol. 9, pp.391–425). Springer-Verlag. [https://doi.org/10.1007/7854\\_2011\\_160](https://doi.org/10.1007/7854_2011_160)
- Koch, C. (2004). *The quest for consciousness: A neurobiological approach*. Roberts & Company.
- Kovacs, P., Hélie, S., Tran, A. N., & Ashby, F. G. (2021). A neurocomputational theory of how rule-guided behaviors become automatic. *Psychological Review*, 128(3), 488–508. <https://doi.org/10.1037/rev0000271>
- Krebs, D., Denton, K., & Higgins, N. C. (1988). On the evolution of self-knowledge and self-deception. In *Sociobiological perspectives on human development* (pp. 103–139). Springer.
- Laughlin, R. B., & Pines, D. (2000, January 4). The theory of everything. *Proceedings of the National Academy of Sciences of the United States of America*, 97(1), 28–31. <https://doi.org/10.1073/pnas.97.1.28>
- Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8(4), 529–539. <https://doi.org/10.1017/S0140525X00044903>
- Libet, B. (1999). Do we have free will? *Journal of Consciousness Studies*, 6(8–9), 47–57.
- Littrell, J. (2008). The mind-body connection: Not just a theory anymore. *Social Work in Health Care*, 46(4), 17–37. [https://doi.org/10.1300/J010v46n04\\_02](https://doi.org/10.1300/J010v46n04_02)
- Loeb, J. (1918). *Forced movements, tropisms, and animal conduct* (Vol. 1). JB Lippincott.
- Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes*, 28(1), 3–26.
- Luke, D. P., & Terhune, D. B. (2013). The induction of synaesthesia with chemical agents: a systematic review. *Frontiers in Psychology*, 4, 753. <https://doi.org/10.3389/fpsyg.2013.00753>
- Mackinnon, D. P. (2011, November). Integrating mediators and moderators in research design. *Research on Social Work Practice*, 21(6), 675–681. <https://doi.org/10.1177/1049731511414148>
- Mansell, W. (2024). An integrative control theory perspective on consciousness. *Psychological Review*, 131(1), 1–17. <https://doi.org/10.1037/rev0000384>
- Marken, R. S., & Mansell, W. (2013). Perceptual control as a unifying concept in psychology. *Review of General Psychology*, 17(2), 190–195. <https://doi.org/10.1037/a0032933>
- Marr, M. J. (2007). Preface: Bridging two cultures. *Behavior and Philosophy*, 35, vii–viii.
- Maturana, H. R., & Varela, F. J. (1991). *Autopoiesis and cognition: The realization of the living* (Vol. 42). Springer Science & Business Media.
- Maugham, W. S. (1915). *Of human bondage*. George H. Doran Company.
- Maxwell, J. C. (1882). Essays. In L. Campbell & W. Garnett (Eds.), *The life of James Clerk Maxwell: With a selection from his correspondence and essays*. MacMillan and Company.
- May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature*, 261(5560), 459–467. <https://doi.org/10.1038/261459a0>
- Mayr, E. (1992). The idea of teleology. *Journal of the History of Ideas*, 53(1), 117–135. <https://doi.org/10.2307/2709913>
- McCulloch, W. S. (1945). The heterarchy of values determined by the topology of nervous nets. *The Bulletin of Mathematical Biophysics*, 7(4), 227–227. <https://doi.org/10.1007/BF02478429>
- McFarland, D. J. (1971). *Feedback mechanisms in animal behaviour*. Academic Press.

- McNamee, D., & Wolpert, D. M. (2019, May 1). Internal models in biological control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1), 339–364. <https://doi.org/10.1146/annurev-control-060117-105206>
- Mechner, F. (1995). *Learning and practicing skilled performance* (pp. 39–42). The Mechner Foundation.
- Metzinger, T. (2013). The myth of cognitive agency: Subpersonal thinking as a cyclically recurring loss of mental autonomy [hypothesis and theory]. *Frontiers in Psychology*, 4, 931. <https://doi.org/10.3389/fpsyg.2013.00931>
- Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. Holt Rinehart and Winston.
- Mitchell, K. J. (2023). *Free agents: How evolution gave us free will*. Princeton University Press.
- Monod, J. (1974). On chance and necessity. In F. J. Ayala, & T. Dobzhansky (Eds.), *Studies in the Philosophy of Biology*. London: Palgrave. [https://doi.org/10.1007/978-1-349-01892-5\\_20](https://doi.org/10.1007/978-1-349-01892-5_20)
- Moore, J. (1981). On mentalism, methodological behaviorism, and radical behaviorism. *Behaviorism*, 9(1), 55–77.
- Moore, J. (2008). Conceptual foundations of radical behaviorism.
- Morris, E. K. (1982). Some relationships between interbehavioral psychology and radical behaviorism. *Behaviorism*, 10(2), 187–216.
- Nadel, L., & Maurer, A. P. (2020). Recalling Lashley and reconsolidating Hebb. *Hippocampus*, 30(8), 776–793. <https://doi.org/10.1002/hipo.23027>
- Nakamura, J., & Csikszentmihalyi, M. (2001). The concept of flow. In *Handbook of positive psychology* (pp. 89–105). The Netherlands, Dodrecht: Springer.
- Négyessy, L., Nepusz, T., Kocsis, L., & Bazsó, F. (2006). Prediction of the main cortical areas and connections involved in the tactile function of the visual cortex by network analysis. *The European Journal of Neuroscience*, 23(7), 1919–1930. <https://doi.org/10.1111/j.1460-9568.2006.04678.x>
- Neuringer, A., & Jensen, G. (2010). Operant variability and voluntary action. *Psychological Review*, 117(3), 972–993. <https://doi.org/10.1037/a0019499>
- Nieder, A., Wagener, L., & Rinnert, P. (2020, September 25). A neural correlate of sensory consciousness in a corvid bird. *Science*, 369(6511), 1626–1629. <https://doi.org/10.1126/science.abb1447>
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know – Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259. <https://doi.org/10.1037/0033-295X.84.3.231>
- Noble, D. (2008). Claude Bernard, the first systems biologist, and the future of physiology. *Experimental Physiology*, 93(1), 16–26. <https://doi.org/10.1113/expphysiol.2007.038695>
- Norman, D. A. (1981). Categorization of action slips. *Psychological Review*, 88(1), 1–15. <https://doi.org/10.1037/0033-295X.88.1.1>
- Parlog, A., Schlüter, D., & Dunay, I. R. (2015, March). *Toxoplasma gondii*-induced neuronal alterations. *Parasite Immunology*, 37(3), 159–170. <https://doi.org/10.1111/pim.12157>
- Parr, R., & Russell, S. (1997). Reinforcement learning with hierarchies of machines. In *Advances in neural information processing systems* (Vol. 10). [https://proceedings.neurips.cc/paper\\_files/paper/1997/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/1997/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf)
- Pearl, J. (2009). *Causality*. Cambridge University Press.
- Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

- Piaget, J. (1952). The second stage: The first acquired adaptations and the primary circular reaction. In M. Cook (Trans.), *The origins of intelligence in children* (pp. 47–143). W. W. Norton & Co. <https://doi.org/10.1037/11494-003>
- Powers, W. T. (1973). *Behavior: The control of perception*. Chicago: Aldine.
- Powers, W. T. (1978). Quantitative analysis of purposive systems: Some spadework at the foundations of scientific psychology. *Psychological Review*, 85(5), 417–435. <https://doi.org/10.1037/0033-295X.85.5.417>
- Prescott, T. J., Diamond, M. E., & Wing, A. M. (2011, November 12). Active touch sensing. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 366(1581), 2989–2995. <https://doi.org/10.1098/rstb.2011.0167>
- Proffitt, D. R. (2006). Embodied perception and the economy of action. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 1(2), 110–122. <https://doi.org/10.1111/j.1745-6916.2006.00008.x>
- Rachlin, H. (1992, November). Teleological behaviorism. *The American Psychologist*, 47(11), 1371–1382. <https://doi.org/10.1037//0003-066x.47.11.1371>
- Rachlin, H. (1998). Teleological behaviorism. In W. O'Donohue & R. Kitchener (Eds.), *Handbook of behaviorism* (pp. 195–215). Academic Press.
- Rachlin, H. (2014). *The escape of the mind*. Oxford University Press.
- Raichle, M. E. (2015). The brain's default mode network. *Annual Review of Neuroscience*, 38(1), 433–447. <https://doi.org/10.1146/annurev-neuro-071013-014030>
- Rescorla, R. A. (1987). A Pavlovian analysis of goal-directed behavior. *American Psychologist*, 42(2), 119–129. <https://doi.org/10.1037/0003-066X.42.2.119>
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.
- Rosenbaum, D. A., Cohen, R. G., Jax, S. A., Weiss, D. J., & van der Wel, R. (2007). The problem of serial order in behavior: Lashley's legacy. *Human Movement Science*, 26(4), 525–554. <https://doi.org/10.1016/j.humov.2007.04.001>
- Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, 10(1), 18–24. <https://doi.org/10.1086/286788>
- Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (2013). Backpropagation: The basic theory. In *Backpropagation* (pp. 1–34). Psychology Press.
- Sachs, R. G. (1987). *The physics of time reversal*. University of Chicago Press.
- Sapolsky, R. M. (2023). *Determined: Life without free will*. Random House.
- Schulte, P. (2021). The nature of perceptual constancies. *Philosophy and Phenomenological Research*, 103(1), 3–20. <https://doi.org/10.1111/phpr.12693>
- Scott, G., Shavlik, J., & Ray, W. (1991). Refining PID controllers using neural networks. In *Advances in neural information processing systems* (Vol. 4). <https://proceedings.neurips.cc/paper/1991/file/285e19f20beded7d215102b49d5c09a0-Paper.pdf>
- Selfridge, O. G. (1958). Pandemonium: A paradigm for learning. In *Proc. Symposium on Mechanisation of Thought Processes* (pp. 513–526).
- Seth, A. (2021). *Being you: A new science of consciousness*. Penguin.
- Shapiro, L. (2010). *Embodied cognition*. Routledge Press.
- Shapiro, L. (2014). *The Routledge handbook of embodied cognition*. Routledge.
- Shepard, R. N. (1978). The mental image. *American Psychologist*, 33(2), 125–137. <https://doi.org/10.1037/0003-066X.33.2.125>
- Shepard, R. N. (1981/2017). Psychophysical complementarity. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 279–341). Routledge.

- Shepard, R. N. (1984, October). Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychological Review*, 91(4), 417–447. <https://doi.org/10.1037/0033-295X.91.4.417>
- Shepard, R. N. (1987). Evolution of a mesh between principles of the mind and regularities of the world. In J. Dupré (Ed.), *The latest on the best: Essays on evolution and optimality* (pp. 251–275). MIT Press/Bradford Books.
- Shepard, R. N. (2008, January 2). The step to rationality: The efficacy of thought experiments in science, ethics, and free will. *Cognitive Science*, 32(1), 3–35. <https://doi.org/10.1080/03640210701801917>
- Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6), 467–482. <https://goo.gl/nQue2P>
- Skinner, B. F. (1969). *Contingencies of reinforcement: A theoretical analysis*. Appleton-Century-Crofts.
- Skinner, B. F. (1988). Responses to commentaries. In A. C. Catania & S. Harnad (Eds.), *The selection of behavior: The operant behaviorism of B. F. Skinner*. Cambridge University Press.
- Skinner, B. F. (1989). The origins of cognitive thought. *American Psychologist*, 44(1), 13–18. <https://doi.org/10.1037/0003-066X.44.1.13>
- Skora, L. I., Yeomans, M. R., Crombag, H. S., & Scott, R. B. (2021). Evidence that instrumental conditioning requires conscious awareness in humans. *Cognition*, 208, 104546. <https://doi.org/10.1016/j.cognition.2020.104546>
- Sosa, R., & Alcalá, E. (2022). The nervous system as a solution for implementing closed negative feedback control loops. *Journal of the Experimental Analysis of Behavior*, 117(3), 279–300. <https://doi.org/10.1002/jeab.736>
- Sperry, R. W. (1966). Mind, brain, and humanist values. *Bulletin of the Atomic Scientists*, 22(7), 2–6. <https://doi.org/10.1080/00963402.1966.11454956>
- Staddon, J. E. R. (2013). *Limits to action: The allocation of individual behavior*. Academic Press.
- Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology & Behavior*, 106(1), 5–15. <https://doi.org/10.1016/j.physbeh.2011.06.004>
- Stewart, I. (2017). RFT as a functional analytic approach to attitudes, beliefs, and motivation. *The Behavior Analyst*.
- Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 1, 444–459.
- Stratton, G. M. (1897). Vision without inversion of the retinal image. *Psychological Review*, 4(4), 341–360. <https://doi.org/10.1037/h0075482>
- Strevens, M. (2011). Probability out of determinism. In C. Beisbart & S. Hartman (Eds.), *Probabilities in physics* (Vol. 146, pp. 339–364). Oxford University Press.
- Sutton, J., McIlwain, D., Christensen, W., & Geeves, A. (2011). Applying intelligence to the reflexes: Embodied skills and habits between Dreyfus and Descartes. *Journal of the British Society for Phenomenology*, 42(1), 78–103. <https://doi.org/10.1080/00071773.2011.11006732>
- Taylor, S. E., Kemeny, M. E., Reed, G. M., Bower, J. E., & Gruenewald, T. L. (2000). Psychological resources, positive illusions, and health. *The American Psychologist*, 55(1), 99–109. <https://doi.org/10.1037/0003-066x.55.1.99>
- Teitelbaum, P. (1966). The use of operant methods in the assessment and control of motivational states. In W. K. Honig (Ed.), *Operant behavior: Areas of research and application* (pp. 565–608). Appleton-Century-Crofts.

- Thorndike, E. L. (1898). Animal Intelligence: An experimental study of the associated processes in animals. *The Psychological Review: Monograph Supplements*, 2(4), i–109. <https://doi.org/10.1037/h0092987>
- Thorndike, E. L. (1927). The law of effect. *The American Journal of Psychology*, 39(1/4), 212–222. <https://doi.org/10.2307/1415413>
- Timberlake, W. (1988). The behavior of organisms: Purposive behavior as a type of reflex. *Journal of the Experimental Analysis of Behavior*, 50(2), 305–317. <https://doi.org/10.1901/jeab.1988.50-305>
- Timberlake, W. (1994, December). Behavior systems, associationism, and Pavlovian conditioning. *Psychonomic Bulletin & Review*, 1(4), 405–420. <https://doi.org/10.3758/BF03210945>
- Timberlake, W., & Allison, J. (1974). Response deprivation: An empirical approach to instrumental performance. *Psychological Review*, 81(2), 146–164. <https://doi.org/10.1037/h0036101>
- Tolman, E. C. (1932). *Purposive behavior in animals and men*. University of California Press.
- Tomasello, M., & Farrar, M. J. (1986, December). Joint attention and early language. *Child Development*, 57(6), 1454–1463. <https://doi.org/10.2307/1130423>
- Tonneau, F. (2004). Consciousness outside the head. *Behavior and Philosophy*, 32, 97–124.
- Turkkan, J. S. (1989). Classical conditioning beyond the reflex: The new hegemony. *Behavioral and Brain Sciences*, 12(1), 121–137. <https://doi.org/10.1017/S0140525X00024572>
- Tyebji, S., Seizova, S., Hannan, A. J., & Tonkin, C. J. (2019, January). Toxoplasmosis: A pathway to neuropsychiatric disorders. *Neuroscience and Biobehavioral Reviews*, 96, 72–92. <https://doi.org/10.1016/j.neubiorev.2018.11.012>
- Vancouver, J. B., & Putka, D. J. (2000, July). Analyzing goal-striving processes and a test of the generalizability of perceptual control theory. *Organizational Behavior and Human Decision Processes*, 82(2), 334–362. <https://doi.org/10.1006/obhd.2000.2901>
- Varela, F. J., Thompson, E., & Rosch, E. (1993). *The embodied mind: Cognitive science and human experience*. MIT Press.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30). [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)
- Waldmann, M. R., Hagmayer, Y., & Blaisdell, A. P. (2006). Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science*, 15(6), 307–311. <https://doi.org/10.1126/science.1121872>
- Werbos. (1988). Backpropagation: Past and future. IEEE 1988 International Conference on Neural Networks.
- Wikipedia. (2024). Logistic map. In *Wikipedia, the free encyclopedia*. [https://en.wikipedia.org/w/index.php?title=Logistic\\_map&oldid=1217803071](https://en.wikipedia.org/w/index.php?title=Logistic_map&oldid=1217803071)
- Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4), 625–636. <https://doi.org/10.3758/bf03196322>
- Zettle, R. D., Barnes-Holmes, D., Hayes, S. C., & Biglan, A. (2016). *The Wiley handbook of contextual behavioral science*. John Wiley & Sons.
- Zukav, G. (2012). The dancing Wu Li masters: An overview of the new physics.
- Zuriff, G. (1985). *Behaviorism: A conceptual reconstruction*. Columbia University Press.
- Zuriff, G. E. (1975). Where is the agent in behavior? *Behaviorism*, 3(1), 1–21. <https://www.jstor.org/stable/27758827>