Assessing Machine Volition: An Ordinal Scale for Rating Artificial and Natural Systems

Running Title: Scale for Measurement of Volition

George L. Chadderdon

Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405

Corresponding Author:

George L. Chadderdon
Department of Psychological and Brain Sciences
Indiana University
Bloomington, IN 47405
Telephone: (812) 322-6947
Fax: (812) 855-4691
Email: gchadder@indiana.edu

Abstract

Volition, though often poorly defined, is a concept of interest and utility to both philosophers and researchers in artificial intelligence. This paper proposes to define volition, and proposes a functionally-defined, physically-grounded ordinal scale and a procedure by which volition might be measured: a kind of Turing test for volition, but motivated by an explicit analysis of the concept being tested and providing results which are graded, rather than Boolean, so that candidate systems may be ranked according to their degree of volitional endowment. Volition is proposed to be a functional, aggregate property of certain physical systems and is defined as *the capacity for adaptive decision-making*. The scale, similar in scope to Daniel Dennett's *Kinds of Minds* scale, is then outlined, as well as a set of progressive "litmus tests" for determining where a candidate system falls on the scale. Such a formulation may be useful for understanding volition and assessing progress made in engineering intelligent, autonomous artificial organisms.

Key Words: volition; free will; ordinal scale; Turing test; adaptive decision-making; artificial intelligence; artificial life

1 Motivation: Why Inquire Into Volition?

The question of whether an artifact can be endowed with a will is both provocative and long-standing, popularized not only in modern science fiction, but also in the nineteenth century story of Frankenstein (Shelley, 1818), in earlier Jewish folklore concerning animated statues, or golems (Idel, 1990), and even in ancient Greek mythology in the Pygmalion and Prometheus myths (Fuller, 1959). Earlier eras were more inclined to anthropomorphically attribute consciousness and willful action to both animals and inanimate objects, but in our current era we tend to regard these as being the exclusive province of higher vertebrate animals with the clearest evidence being shown by mammals. The emergence of computer technology in the twentieth century, however, has brought with it the increasing possibility of engineering artificial organisms of similar complexity to their natural precursors. In the future, these artifacts may succeed in convincing us that they engage in "freely" willful behavior, i.e., that they possess the attribute of *volition*.

It is primarily the disciplines of artificial intelligence (AI) (Rich & Knight, 1991; Russell & Norvig, 2002) and artificial life (ALife) (Langton, 1989; Levy, 1992) that are concerned with the Promethean project of engineering the capacity for human volition in artifacts. Unfortunately, although much progress has been made in the past few decades, the science-fiction-envisioned goal of android intelligence seems distant still. Three factors that may be impeding progress in AI include: a) a methodological tendency to design systems that implement particular domains of expertise in human intelligence, but fail to generalize across domains (Dreyfus, 1979); b) a need to leverage the vast, ever-accumulating knowledge regarding the neural mechanisms of behavior; and c) a frequent focus on engineering of knowledge-based systems which do not exhibit the kind of autonomous behavior we expect from volitional agents. It is primarily the last issue that this paper is concerned with.

Traditional AI, being focused on cognitive information processing, often deemphasizes the situated agent aspect to intelligent behavior. In addition to being an information processor, an organism is also an autonomous actor, a source of behaviors that influence the environment which, in turn, influences the agent: both its immediate perceptions and its future reactions to the environment (Weng et al., 2001).

This is the essence of what most people mean when they consider the concept of willful activity. Robert Kane proposes that an agent possesses volition—he uses the term *free will*—when it bears the "ultimate responsibility" for its acts (Kane, 1996). Although this notion of ultimate responsibility is ultimately problematic (Dennett, 2003), it does capture the intuition that volitional agents are autonomous actors whose decisions and behaviors play a key role in determining their fates.

In the view of this paper, engineering agent-hood in artificial systems should be a primary goal of AI and ALife. Embodied cognitive approaches to AI (Anderson, 2005; Pfeifer & Scheier, 1999) are generally on the right track because of their focus on engineering whole autonomous agents. Generally, we expect intelligent beings to possess volition. Systems need to exhibit signs of animacy and goal-directedness in order to be regarded as volitional. Humans either judge or perceive intention and animacy of systems based on visual motion cues (Barrett, Todd, Miller, & Blythe, 2005; Scholl & Tremoulet, 2000), so systems that only perform disembodied information processing become unconvincing exemplars of intelligence. In the author's view, moreover, not only is volition necessary for intelligence, but what we call intelligent behavior emerges from systems that possess a sufficient degree of volition.

That volition should be regarded as a graded rather than an all-or-nothing phenomenon is the fundamental argument of this paper, and its main task is to propose a rough definition and a scale of measurement for volition, including a set of empirical questions one might ask in order to determine where a natural or artificial system lies on that scale. Whereas Turing's test (Turing, 1950) sought to answer through a single empirical "litmus test" whether a system was intelligent, this paper proposes a set of such tests to rank systems along an ordinal scale, providing an additional advantage of giving a conceptual analysis of the concept being measured, i.e., volition.

But why should researchers inquire about volition at all? AI researchers might simply take the Turing approach and conclude that if they've created a system that acts the way humans appear to, they've succeeded in AI's mission, and any consideration of whether a system possesses or lacks volition is esoteric. Cognitive scientists might decide to simply analyze the brain's mechanisms of behavior and not worry about which aspects of behavior count as being volitional.

This paper argues that volition is a useful concept to analyze and measure in systems because it may suggest both a set of distinct architectural features and a plan of attack for researchers who are trying to understand the mechanisms of motivated human behavior. Setting a single criterion for volition in the manner of Turing's original test may be useful, but it is a blunt instrument of analysis. What both cognitive scientists and AI researchers are most concerned with are the *mechanisms* of willful, intelligent behavior: the architectural features and subsystems that allow the complex phenomena of volition, intelligence, and consciousness to be manifest in physical systems. Volition, like intelligence and consciousness, seems to be possessed in different degrees by organisms, and the difference of degree may be best explained by which functional and neural architectural features the organisms possess or lack. More primitive animals will possess the most basic of these features, whereas more complex and flexibly-behaving animals such as primates and humans will possess the more advanced features.

Once a set of functional architectural features can be isolated in natural and artificial systems and these are ordered, as best as they may be, along a continuum, the resulting scale may be employed for a number of useful purposes. Cognitive neuroscientists may attempt to find which brain mechanisms correspond with the functional features, allowing the problem of the neural substrate of volition to be attacked piecemeal. AI researchers may benefit from the discoveries of cognitive neuroscientists by designing implementations and algorithms that more closely match the working functionality in animal brains, or they may attempt to develop their own mechanisms that fulfill the functional specifications of the different scale levels. Either way, the scale may usefully break down the problem of volition into more manageable components. The ordering of these components along the scale may suggest an ordering for AI researchers in developing the mechanisms of the specific volitional systems. Similar to the layered design and implementation approach of Rodney Brooks (1991) or Aaron Sloman (2003), mechanisms might be developed for the more primitive architectural features and, once these primitive layers functioned effectively, the more abstract or high-level layers could be developed around them. This type of design has some precedence in natural organisms because mammalian brains possess a layered structure with more primitive, evolutionarily older brain structures deeper in the interior of the

brain regulating critical homeostatic processes, and newer, more exterior brain structures adding increasing behavioral flexibility that is needed for evolutionarily more recent organisms (MacLean, 1990).

Besides directing the attention of researchers to component mechanisms, however, a scale for volition may serve as an evaluative tool. If a set of empirical tests can be specified for each level of the scale, both natural and artificial organisms may be ranked and compared for volitional endowment. While some allowance must be made for the essential multidimensionality of complex phenomena such as volition (see supplementary material, Section S3.3), an ordinal scale of volition would provide a means of comparing artificial systems with natural ones and with each other, allowing AI to better gauge its progress at engineering convincingly intelligent and volitional systems.

[<--- Insert Figure 1 near here. --->]

The remainder of the paper, laid out in three sections, will address the ontological problem of defining and analyzing volition (Section 2), the epistemological problem of measuring it in natural and artificial systems (Section 3), and an examination of how the resulting scale might provide insights towards the goal of engineering systems with a human degree of volition (Section 4). To summarize, volition will be analyzed as a linear scale consisting of four levels, each of which consists of a set of two or more sublevels. Each succeeding level in the hierarchy includes features of all of the earlier levels, and corresponding to each level, there is a "litmus test" by which a researcher may empirically decide whether the system in question falls into the level (provided it has passed all of the tests of the previous levels). Figure 1 shows an overview of the scale, and Tables 1-4 show more detailed information for each level of the scale, including all of the "litmus tests" for the sublevels. Supplementary material, available online at www.indiana.edu/~cortex/volitionscale/supplemtext.pdf [...and with the journal's online text at the Sage site…], provides additional details related to the sublevels (Section S1), examples of use of the scale (Section S2), and a discussion of potential objections and issues with its current conception (Section S3).

The approach of analyzing complex mental concepts by scales based on accumulated architectural or functional features has a number of precedents in the literature. Valentino Braitenberg in

Vehicles: Experiments in Synthetic Psychology (1984) proposes a hierarchy of structural and functional design features that would generate volitional behavior in small, wheeled robots. Rodney Brooks' subsumption architecture, which has been used for developing several robots, is built of nested reactive, hard-wired layers (Brooks, 1991). More recently, Sloman (2003) has proposed a three-layer structure for emotional processing in artificial systems. This was, in part, inspired by Daniel Dennett's *Kinds of Minds* (1996) which proposes a broad-brush four-level evolutionary scale of mind: his Tower of Generate-and-Test. Verschure and colleagues (Verschure, Voegtlin, & Douglas, 2003) have proposed and implemented robots using a three-layer architecture consisting of reactive control, adaptive control, and contextual control layers. This architecture has some analogous features to those found in animal brains (see Section 4).

2 The Ontological Question: What is Volition?

In seeking to define a concept, it seems sensible to begin with an analysis of common usage. Here, it is especially appropriate because whether or not an artificial system is perceived as being volitional will depend upon whether the system's behavior is at least consistent with this usage. Webster's New World Dictionary ("Webster's New World Dictionary & Thesaurus", 1998) gives three definitions for 'volition'. The most relevant for our purposes is: "the power or faculty of using the will." The most relevant definition of 'will' in the same dictionary is: "the power of making a reasoned choice or decision or of controlling one's own actions."

Upon consideration of these definitions, we can notice some important aspects of what is delineated as volition. It is a property possessed by some entities in the world: the property of being able to make a choice, i.e., a choice on some action the entity might engage in. This action might be an overt action, like locomotion, orienting sense organs, or emitting a social call; or a covert action, such as shifting attentional focus to color rather than shape features, revising a previous evaluation of a particular object encountered in the environment, or mentally selecting a future overt action out of a set of behavioral candidates. A further component of the definition of 'will', "reasoned choice", suggests that

there is some internal mental representation of reasons: representation which may be thought of as values or goals.

With this common usage in mind, our task in this section will be to iteratively develop a suitable definition for the concept of volition and specify an ontology for operationalizing it. Many definitions can and have been proposed which are generally consistent with common usage, but are either too permissive or restrictive, depending on individual tastes. Conceiving volition as a scale allows both the permissive and restrictive usages to be accommodated into one concept consistent with common usage. The remainder of this section, after making explicit the underlying metaphysical assumptions and briefly discussing the ontological status, i.e., the "reality", of volition, develops through successive attempts the finalized formulation of the definition, including the individual levels and sublevels of the scale.

2.1 Metaphysical Assumptions

The approach in this paper assumes a functionalist materialist answer to the mind/body problem. [1] For most AI researchers, this is unproblematic, for the hope of successfully reproducing human intelligence (or volition) in machines rests on the idea that if a physical system is organized in a sufficiently isomorphic way to natural intelligent systems, it will exhibit "real" human thought, not just a clever simulation thereof.

Another relevant metaphysical question is the relationship of volition to determinism. The author's own position is that of a *compatibilist*, i.e., that there is a concept of volition, or "free will", which can exist in a deterministic universe, and which meets all of the concerns we care about (Dennett, 1984, 2003). The term 'volition' seems preferable to "free will", however, because discussions of free will carry a lot of historical metaphysical baggage and "free will" is easily construed as meaning *agent-causation* which is the doctrine that we are "prime movers unmoved", that "In doing what we do, we cause certain events to happen, and nothing—or no one—causes us to cause those events to happen." (Chisholm, 1964). Such a view does not encourage inquiry into causal mechanisms of free choice, and it leads to unrealistic ideas of what freedom must mean. 'Volition' is a more metaphysically neutral term, so it will be used here instead of "free will".

That said, the proposed scale and methodology doesn't make any critical assumption about determinism vs. indeterminism. Kane (1996) provides a plausible story for how a naturalistic, causal, indeterministic system might produce choosing behavior, though his arguments that the mechanisms of human choice must be indeterministic seem questionable (Dennett, 2003). Either deterministic or indeterministic causal mechanisms could be developed for the various levels of volition that will be proposed. Deterministic and probabilistic approaches to implementation both seem viable theoretically, though implementation may favor that indeterministic approaches be approximated with chaotic pseudo-randomness.

2.2 Ontological Status of Volition and a Preliminary Formulation

There is much debate in the philosophy of mind literature on whether propositional attitudes (PAs)—i.e., "folk psychological" constructs such as desires, intentions, and beliefs—really exist or whether they are vacuous concepts such as phlogiston or ether were in chemistry and physics (Churchland, 1981; Clark, 2001; Dennett, 1991; Fodor, 1987). This is relevant to our discussion of volition because volition is closely related to intentions and to desires. If intentions and desires do not have real ontological status, then neither does volition which can be thought of as a faculty that operates on intentions and desires.

This author takes the position elaborated in Dennett's *Real Patterns* paper (1991) which regards PAs (and volition) as having the same kind of ontological status as other *abstracta* such as centers of gravity or field-lines of electromagnetic force. Electromagnetic field-lines do not exist in the literal sense that there really are lines around a magnet or an electric charge, but they do exist in the sense that the physical patterns which cause us to explain electromagnetism in terms of force-lines are real. E-M field-lines can therefore be regarded as real even though, precisely speaking, they may not have the same kind of physical existence as tables and trees.

In Dennett's writings (1991, 2003), he uses the example of Conway's Game of Life to illustrate the ontological status of volition and to demonstrate how volition might be possible in a deterministic universe. Briefly, the laws of a cellular automaton may be entirely deterministic, but if you construct a

sufficiently complex Game of Life configuration, such that the gliders, blinkers, etc., implemented a Turing machine programmed to play a game of chess, then the system would exhibit intentions, i.e., by choosing the next chess moves. You would not be able to detect the goal-orientedness of the system by looking at the cellular automata rules, nor by watching individual gliders, etc. It is only by assuming what Dennett calls an *intentional stance* towards the system, that one could perceive that it is, in fact, a teleological system. Similarly, in biology a teleological approach to the study of an animal's behavior would involve adopting an intentional stance, since it is too difficult to derive an organism's goal-directedness from phenomena at the molecular or cellular level.

Thus, volition can be said to have real ontological existence even though it may not be apparent in the cellular functioning of an organism. We can say that volition in the chess system is real precisely because the system engages in choosing behavior. Volition is an emergent collective property that describes dynamics at a macro-level of existence, i.e., the level of an organism engaged in behavior and decision-making. In this view, volition can be viewed as nothing more than executing one behavior rather than another in a situation according to some (implicit or explicit) goal. What makes the behavior a choice rather than merely a regular, fixed consequence is that in *similar* states of the environment, the organism may well engage in a different behavior.[2,3]

This formulation effectively captures the notion of volition as synonymous with 'autonomy', i.e., self-motivated action, because differential behavior in the face of similar environmental conditions implies that some difference of internal 'self' state was necessary to account for the difference in behavior. This internal state is effectively 'owned' by the agent/system, consistent with the intuition of a volitional system having a degree of "ultimate responsibility" (Kane, 1996) for its acts.

At this point, it may seem that we've arrived at a satisfactory operational definition of volition: the capability of exhibiting a different behavior in a sufficiently similar environmental context. However, if we accept this without refinement, we must consider thermostats and video-game "bots" to be in the same volitional class as humans who have capacity for verbal deliberation and planning and reflection.

That this seems troublingly inadequate suggests that we might want to rethink considering volition as a Boolean property.

2.3 Volition as an Ordinal Scale

At its most basic level, volition amounts to the property of being a self-motivated actor in the world, possessing some ongoing autonomous behavior, where 'autonomous' means that the behavior of the system is best explained by internal rather than external factors (e.g. a clock's change of display, as opposed to a rock's rolling down a hill). But this seems too permissive a criterion; it eliminates rocks and other inert objects, but allows clocks, motors, and even conceivably bodies of water to be counted as possessing volition.

At the most refined level, volition implies not only basic autonomy, but also having the ability to use internal verbalization to ponder questions and plan courses of action in advance and to deliberate on whether particular choices are likely to lead to positive or negative outcomes before committing to a particular action. Surely, however, this criterion is too restrictive. Even non-primate mammals lacking language may still exercise some kind of nonverbal deliberation process when, for example, foraging, or finding their way back to their lairs. A case for volition could be made even for insects with their largely reflex-driven behavior.

Therefore, it seems that volition may be better conceived as a graded property rather than as a dichotomous Boolean property. Lacking volition entirely are inert objects like rocks and cups and spoons. There are also entities that spontaneously enact some sort of behaviors like clocks and rivers and suns, but cannot be said to possess goals, so they may be disqualified as well. On the low end of volitional endowment are some artifacts like thermostats and heat-seeking missiles that do appear to exhibit some (implicitly) goal-seeking behavior. On the high end are beings such as humans that plan and make predictions and use language and cultural artifacts as aids in directing decision-making processes before committing to an action. In between is the whole animal kingdom and some of mankind's more cunning technological creations. Consideration of volition as falling along a scale may permit a better

systematic understanding of the concept and, additionally, might allow us to gauge the degree to which particular individuals or classes of artificial and natural organisms possess volition.

Before laying out a scale, though, we should ask what general dimension the scale is moving along, what feature is increasing in amplitude as we ascend. At the beginning of Section 2, we noted that common usage suggests that volition involves the ability to choose actions based on values or goals. Another way of stating this is to say that volition entails the capacity for adaptive decision-making. Decision-making implies that differences in the environment and one's internal state (e.g. degree of hunger or fatigue) select behavior. *Adaptive* decision-making implies selection of behaviors that are conducive to meeting goals which will either be those of approach or of avoidance.[4] In natural systems, goal-seeking behavior (e.g. finding food and mates, and escaping from predators) is important for the survival and propagation of the organism; this explains how goals may be said to have evolved in natural organisms. We now have a useful, naturalistic general definition of our concept:

Definition 1. *Volition is the capacity for adaptive decision-making.*

An increase in the capacity for adaptive decision-making, then, would mean that the system would effectively control more of the processes for choosing one behavior vs. the other possible behaviors, and as a result of the increased flexibility, the adaptivity, i.e., the beneficial effect, of its behaviors would probably increase when the organism was in a complex, dynamic environment. Boden (1996) makes relevant observations regarding what is necessary for an increase of volition—she uses the term 'autonomy'. It is said to be greater in organisms to the extent that: 1) their behavior is guided more by learned experience as opposed to "hard-wired" reflexes, 2) their controlling mechanisms emerge more through a development process as opposed to being fixed at the beginning of their existence, and 3) their controlling mechanisms can be more regularly reflected upon and self-modified.

As volition increases, systems can be observed to have behavior differentiated on increasingly subtle variances in the environment and internal state, and their behavioral choices also become more frequently effective in meeting their goals, both learned and innate. Several factors are included in this capacity: increasing sensitivity to subtle distinctions in the environment, higher-order integration of

context-insensitive percepts into more context-sensitive concepts, increasing sensitivity to temporal patterns allowing predictions, increasing self-perception and modeling of internal state, and increasing sphere of influence on the system's environment, expanding both spatially and temporally. As organisms gain in volition, their ability to make self-advancing choices is increased and so is their effect on their environment. An ant acts reflexively and has a small, narrowly circumscribed, and short-lived sphere of influence. An encultured modern human, by comparison, has a sphere of influence extending potentially over the entire globe, and for decades in time.

For any scale, it is generally necessary to provide a unit to measure it with. However, it is not apparent what real number or even integral measure might be useful for volition. The degree of volition possessed by the system seems, rather, to depend on what functional or neural mechanisms it is endowed with. If the set of the relevant functional mechanisms or attributes could be elaborated and ordered, the result would be an ordinal scale: a set of levels of increasing volition. Ascending a level would mean acquiring a fundamental new property that increases flexibility and adaptivity of behavior. One might imagine an engineer developing a volitional system starting with an inanimate object and then adding in new features one by one to create a more flexible and adaptive system. The scale proposed in this paper enumerates a set of such functional architectural features and provide a suitable ordering for them, so that as we add in each new feature to a system, it can be viewed as gaining in volition. As shown in Figure 1 and Tables 1 through 4, the current conception of this scale is composed of four main levels, each broken down into sublevels. The remainder of this section describes the main levels in more detail. More details regarding the individual sublevels are given in the supplementary material in Section S1.

[<--- Insert Table 1 near here. --->]

2.3.1 Level 0: Non-Volitional Systems

Objects and systems in this level can be said to lack volition, either because they engage in no autonomous behavior or because their behavior is not driven by mechanisms that could be described as being goal-directed. Not all non-volitional objects are alike, however. Some seem to exhibit more spontaneous behavior than others, and it is worth pausing to see what features get added along the way

between completely *inanimate objects* (Level 0.0) and automata that achieve some degree of volition at Level 1. *Schizoid automata* (Level 0.1) exhibit self-determined behavior, but are unresponsive to environmental cues. *Reactive automata* (Level 0.2) add stimulus-dependent influence to this capacity for internally-driven behavior. Although they do not possess a goal-directedness that would allow us to consider them as being volitional, it could be said that Levels 0.1 and 0.2 exhibit a kind of *proto-volition* because of their autonomous behavior. Table 1 gives a summary of Level 0's sublevels and Section S1.1 discusses them in more detail.

[<--- Insert Table 2 near here. --->]

2.3.2 Level 1: Instinct-Driven Automata

The first (macro-) level of systems possessing volition consists of automata that are driven solely by hard-wired mechanisms (which we may call *instincts*), but can be said to possess rudimentary goals/values in the form of reflexive tropisms. This level corresponds to Dennett's *Darwinian creatures* level in his Tower of Generate-and-Test scale (Dennett, 1996). *Value-driven automata* (Level 1.0) have a single fixed tropism or set of concurrent, continuously-active tropisms, whereas *modal value-driven automata* (Level 1.1) possess multiple tropisms whose current activation, however, is dependent on either environmental cues or internal state. It is possible to attribute some degree of *choice* to systems at Level 1.1 since they may behave in noticeably differing ways under similar environmental contexts. Table 2 shows a summary of sublevels for Level 1, and Section S1.2 discusses them in more detail.

[<--- Insert Table 3 near here. --->]

2.3.3 Level 2: Contained Self Organisms

The next level of volition is reserved for organisms that rely chiefly on internal cognitive mechanisms (their "naked brains") in order to adapt to their environments; that is, they engage in little of what Clark (1997, 2001) and others have dubbed "scaffolding" on their environments. (It is granted that many animals engage in some sort of at least social scaffolding, but we take most solitary types of wild animals that have little faculties for manipulating objects as falling into this level.) *Associative learning organisms* (Level 2.0) add the capacity for learning and conditioning to the instinctive mechanisms they

possess from Level 1. The boundary between this sublevel and earlier levels may be considered the Great Divide of volition, and it corresponds to the transition between the first (*Darwinian creatures*) and second level (*Skinnerian creatures*) of Dennett's Tower of Generate-and-Test scale (Dennett, 1996).

Moving to the next sublevel, *ideational organisms* (Level 2.1) possess some form of dynamic short-term memory (STM), or *working memory*, which allows them to maintain and release mnemonic contexts which may serve as representations of tasks or parameters of task performance. *Recollective organisms* (Level 2.2) also have declarative long-term memory (LTM) which allows both storage of facts (*semantic memory*) and representations of perceptual experiences (*episodic memory*). Such organisms are capable of drawing upon representations of past experiences in a case-based fashion to select a behavior from their repertoire.

Finally, *deliberative organisms* (Level 2.3) have the ability to coordinate the use of their short- and long-term memory with attentional mechanisms that allow them to deliberate on potential actions in response to the current situation before committing to a response. This last capability is a hallmark of advanced volitional capacity, and a minimal requirement for any android-level intelligence. With Level 2.3, we arrive at Dennett's *Popperian creatures* Tower of Generate-and-Test level (Dennett, 1996).

Table 3 summarizes the sublevels of Level 2, and Section S1.3 discusses them in considerably more detail.

[<--- Insert Table 4 near here. --->]

2.3.4 Level 3: Extended Self Organisms

Much of the animal kingdom probably falls under the domain of Levels 1 and 2. However, in birds to a certain extent, and especially mammals, we begin to see a higher level of volitional functioning emerging, one which seems to depend on the animal's ability to extend its mental processing into its environment. Volitionally lower animals, including insects, build nests, dig burrows, and engage in complex social behaviors, but these lack features of volition encountered at Level 2 (with the possible exception of reinforcement learning). They mostly give the impression of being hard-wired and stimulus-driven in their adult behavior. Even some mammals, i.e., the solitary kind that lack impressive object-

manipulative skills, may be considered to be at a volitional deficit when compared with organisms that use the environment to bootstrap their mental processes.

The idea of an *extended mind* is that artifacts and other external entities in the environment might constitute a critical part of an organism's mental processing such that the phenomenon of mind doesn't stop at the "skin-bag" that bounds an organism's body (Clark, 2001). In the view taken in the present paper, extension of mind allows extension of volition because the utilization of external resources allows augmentation of an organism's natural abilities which permits increased adaptivity of behavior. Because the external environment is more rapidly malleable than an organism's genome, extension of mind enables rapid adaptation (in the larger sense, with learning included) in the face of changing conditions. An individual organism can, in fact, share extensions with other conspecifics that improve their chances of survival (as well as the survival of the group or the gene-pool).

Social organisms (Level 3.0) are capable of social signaling and response which allows them to incorporate the social environment of their herd/flock into their mental processes, and indeed, to construct their own social niche. *Manipulative organisms* (Level 3.1) add the capability of manipulation of objects in the environment. Whereas there may be a limited set of social or alarm calls available to Level 3.0 organisms, *symbolic organisms* (Level 3.2) possess a communicative system which has syntactic structure which enables a more flexible representation and communication of information. Finally, *cultural organisms* (Level 3.3) have developed an external means (e.g. writing) to encode their symbolic communications which allows rules and customs to be formally codified and passed down through generations and across long distances.

Table 4 summarizes the sublevels of Level 3 and more details are given in Section S1.4. Dennett's final level of his Tower of Generate-and-Test (*Gregorian creatures*) corresponds roughly with this level, though Dennett places greater emphasis on the role of language (Dennett, 1996).

3 The Epistemological Question: How Do We Measure Volition?

Having designed an ordinal scale for volition, we now hope to be able to measure the volition of prospective systems against it. In Tables 1 through 4 are provided "litmus questions" for each of the levels. To classify a candidate system, one would begin at Level 0.0 and iteratively ascend the scale, answering the litmus questions until the first No answer was encountered. The system would be categorized at the sublevel of the final Yes answer. Thus, membership in a volitional sublevel entails passing all of the litmus tests up to and including that sublevel. Section S2 of the supplementary material provides two examples of this evaluation process: one for Pac-man ghosts (video-game entities), and another for chimpanzees.

3.1 Level 0 (Non-Volitional Systems) Litmus Questions

The two questions at this level are reasonably objective and simple to evaluate. Level 0.1's question is: *Does the system move or act on its own, i.e., without continuous prompting by external forces?* If an object moves in such a way that it can't be explained merely by gravitation or other external forces, then the system passes the test. Otherwise, it fails and the system may be considered inanimate.

The question for Level 0.2 is: *Is the system's spontaneous behavior modified by events/conditions in the environment?* If the system always engages in the same behavior (whether rhythmic or random) and the different conditions in the environment fail to generate different behaviors in the system (that can't be explained by things like gravitation, etc.), then the system fails the test and is a schizoid automaton.

3.2 Level 1 (Instinct-Driven Automata) Litmus Questions

The questions at this level are also fairly straightforward. Level 1.0's question is: *Does the system appear to be trying to approach or avoid any object or occurrence of an event in its environment?* Rocks rolling down a hill towards the bottom of the slope have already been eliminated by the test for Level 0.1, so we needn't worry about them at this point. If the stimulus-receptive system in question shows itself to approach or avoid either physical objects or conditions or events in the environment, it passes this test. Otherwise, it fails and is merely a reactive automaton.

Level 1.1's question is: *Does the system have different sets of goals active during different environmental or bodily conditions?* This test is most obviously passed if different environmental conditions can be correlated with different approach/avoidance behaviors in the system and a double dissociation can be established between the sets of goals followed. If internal state is the decisive factor, then some physiological analysis may need to be done. In a biological animal, for example, correlation of differences in approach/avoidance behavior to high or low levels of blood glucose (which would correlate with hunger, presumably) would constitute a means of passing this test. If the test fails, the system is only a (non-modal) value-driven automaton.

3.3 Level 2 (Contained Self Organisms) Litmus Questions

At this level, the testing questions begin to become more difficult and observation of the external behavior of the system may not suffice in arriving at an answer. Level 2.0's question is: *Does the system develop new adaptive approach or avoidance patterns over time?* This can be determined by periodically placing the system in the same situations it was in previously. If it shows new or modified approach/avoidance behavior, then it passes the test. Otherwise, it is merely a modal value-driven automaton.

Level 2.1's test question is: *Can the system engage in a task that requires working memory (e.g. delayed non-match-to-sample)?* A delayed match- or mismatch-to-sample test is probably the most straightforward way of testing this, or a test analogous to a fear trace conditioning study (Han et al., 2003) where mice were trained to associate a tone with a foot shock (evoking a fear-motivated freezing response) where the shock was delivered after a delay period after the offset of the tone. The system needs to be trained on a task that depends on the ability to maintain the representation of a stimulus in its absence. If it can't be trained on any such task, then the system fails the test and can be considered only an associative learning organism.

Level 2.2's test question is: *Can the system engage in a task that requires long-term memory?* There may be some difficulty determining whether or not a task was solved using long-term or short-term memory or using just reinforcement learning. Short-term memory can be eliminated as a possibility by

having a long delay period in the task. Differentiating long-term memory recall from conditioning using just behavioral cues may be more difficult. With artificial systems, however, we will probably have a better sense of the internal structure and will be able to determine if long-term memory was used. An example of strong evidence in natural systems includes such behaviors as food caching in scrub-jays (Clayton, Emery, & Dickinson, 2006); the birds not only seem to remember the locations where they hide their food, but if food they stored is perishable and a long period of time has passed, they will not visit their likely-spoiled caches.

Level 2.3's test question is: *Can the system engage in a behavior (e.g. game-playing, navigation) that requires evaluation of multiple possibilities without action?* This may also, admittedly, be a difficult question to answer because just working memory in combination with reinforcement learning may be capable of finding cunning non-deliberative solutions to very complicated tasks and it is hard to differentiate deliberative solutions from non-deliberative ones just by external observation of the organism. Experiments such as done with the chimpanzee Julia (see supplementary material, Section S2.2) can provide strong evidence of deliberative behavior by requiring the system to preplan its sequence of actions before it initiates them. For systems that we know to have localized long-term memory, we might do a kind of brain-scan in order to determine the patterns of long-term memory access during ongoing behavior that might benefit from deliberation. If we know localized evaluative centers (such as the amygdala in animals) then we might also determine when evaluation is being done during ongoing behaviors. Obvious patterns of: ACCESS, EVALUATE, ACCESS, EVALUATE, ACT would strongly suggest deliberative behavior is taking place. If no such patterns can be found, then the organism could be considered to be only a recollective organism.

3.4 Level 3 (Extended Self Organisms) Litmus Questions

The questions at this level vary in difficulty, but they have the advantage of not requiring much physiological analysis since they are mainly about patterns of behavioral interaction with the environment. The litmus question for Level 3.0 is: *Does the organism send and selectively respond to social cues?* This sounds potentially difficult, but with animals, it seems to be pretty straightforward. If

an artificial system known to have at least Level 2.3 status engages in signaling that affects the behavior of other systems of the same type, then it passes this test. Otherwise, it is only a deliberative organism. Of course, there are different degrees of social function. Such evidence as herd-formation, dominance hierarchy, and specialized division of labor would provide even better evidence.

The question for Level 3.1 is: *Can the system pick up and move around objects in its environment?* This is easily observable for animals, though for virtual organisms operating in rich, but virtual environments some allowances might need to be made for what constitutes manipulation. Moving a virtual chess-piece might count for a game-playing system, for example.

The question for Level 3.2 is: *Does the system communicate using language that has syntax as well as semantics?* This is rather difficult question, to be sure. There may be much debate even now as to whether cetaceans and other higher mammals might possess the ability to use symbols, though there is evidence that properly-trained chimapanzees such as Kanzi may use noun-verb "sentences" (Deacon, 1997), and even some evidence that parrots may answer questions about different kinds of classes of properties (Pepperberg, 2002). A thorough discussion of what might constitute symbol use, complete with syntax, can be found in (Deacon, 1997). Mere use of threat-selective alarm calls isn't enough: syntax requires that the symbols be combinable in larger informational structures.

Finally, the question for Level 3.3 is: *Does the system use external storage of symbols as a repository for knowledge?* Once we've established symbol usage in an organism, answering this question seems straightforward. If they can store their symbols and pass them along, they pass this test. Otherwise, they are just symbolic organisms.

4 Conclusion: Progressing Towards Human Volition

This paper, clearly, takes the "strong AI" position that machines may one day be endowed with volition, even at the level of humans, provided they are designed and implemented according to the correct functional principles. What these functional principles are and how we may extract them, however, is still an open question. Among AI researchers there has been much variance in methodology regarding the degree to which designed systems mirror the structure of natural systems, in particular how

much attention is paid to the organization and mechanisms of the human brain. While a non-neurobiologically motivated solution to AI may, in principle, be found in the future, consideration of mechanisms of brain function will, arguably, facilitate progress in the field immensely, despite the analytical complexities involved. Although the scale proposed here has made no set assumptions about implementation, including neural substrate in animals, it was designed with the context of mammalian neural organization in mind. We will conclude by using the proposed scale to survey some of the work done in AI and ALife and gauge what advances may be needed in order to achieve the goals outlined in this paper. Some pointers will be offered to the comparative and computational neuroscience literature which may be of interest to researchers who wish to develop mechanisms inspired by what is known about the neural mechanisms of animal cognition and behavior.

According to our scale, Level 1.0 (Value-Driven Automata) is the minimal level for volition and yet a surprising amount of research has not produced systems that achieve this level. Whether symbolic, connectionist, or hybrid approaches are used, expert systems that analyze information and output classifications, diagnoses, etc. do not meet this level because they engage in no goal-seeking behavior. To be fair, much of this research has yielded valuable insights about possible mechanisms for learning and pattern classification, but the systems that result are passive, not active agents. Even board-game-playing systems (e.g. chess and backgammon), which might otherwise be regarded as Level 1.1 (Modal Value-Driven Automata) systems because there is a pursuit of goals that may vary according to the situation on the board, nonetheless could be argued to be Level 0.0 (Inanimate Objects) because they often sit idle until a keystroke or mouse-click is made. The same could be said of ELIZA-like “chat-bots” such as ALICE (which stands for Artificial Linguistic Internet Computer Entity) ("A.L.I.C.E. Artificial Intelligence Foundation", 2008) or Jabberwacky ("Jabberwacky", 2008) (though the face accompanying ALICE conversations has eye-tracking that follows the cursor on the browser). Generally, we expect true autonomous behavior to be continuous, whether another agent is present or not. So the first relevant advance to be suggested for volitional systems is: *Volitional systems should exhibit independent, continuous behavior that is responsive to the environment.* Thinking about the designed systems as

autonomous agents rather than collections of procedures that perform cognitive operations is a first step in encouraging this, a shift of emphasis which has already been made by embodied cognitive approaches.

Much of human technology operates on the negative feedback principle of the humble thermostat, but this suffices for Level 1.0 (Value-Driven Automata). Computer opponents in video-games tend to fall into Levels 1.0 or 1.1 (Modal Value-Driven Automata), depending on how sophisticated the programmer's design of their behavior. Impressive and relatively convincing life-like behavior can be accomplished by modal value-driven automata whose sets of active goals change according to different environmental or internal cues. Some state-of-the art mobile robots, including bipedal walking robots such as Honda's Asimo ("Honda Worldwide | ASIMO", 2008) or UT Munich's "Johnnie" biped (Löffler, Gienger, & Pfeiffer, 2003), fall under the domain of Level 1 because they have a fixed repertoire of behaviors and lack learning. Evolutionary algorithm approaches to AI and ALife (Bäck, 1996) that attempt to evolve behaviors to maximize fitness often result in generation of agents that fall within Level 1 because all of the learning is across generations rather than during the life-cycle of the modeled organisms. Some evolutionary ALife approaches such as PolyWorld (Yaeger, 1993) also incorporate neural plasticity into their architectures, which allows them to advance into Level 2.0 (Associative Learning Organisms). The principle to be observed in the highest Level 1 automata is that: *Volitional systems should have sets of goals that are dependent on the system's internal state or environmental conditions*. Patterns of responsiveness should change according to context, whether external or internal.

Even for organism design at Level 1 (Instinct-Driven Automata), AI researchers could gain considerable insight by studying the neural mechanisms of emotion and motivation that are common to all mammals and much of the animal kingdom in general. There is evidence, for example, that there may be in mammals dedicated subcortical systems for SEEKING, RAGE, FEAR, LUST, CARE, PANIC, and PLAY (Panksepp, 1998). A functional understanding of these systems may give AI researchers a better sense of what the behavioral modes are that select the currently active and operative values of an organism. Evolutionary computational approaches might, for example, adopt evolvable building blocks

that control the operations of these various motivational subsystems and how they interact to generate the basic “temperament” of the evolved individuals.

The frontier of AI appears currently to be at Level 2.0 (Associative Learning Organisms), the level which first incorporates learning. As of 1999, Sony’s Aibo robotic pets have reinforcement learning features that allow their owners to train them through praise and scolding (Sony Corporation, 1999). Some recent embodied cognitive computational neuroscience work has also been done incorporating biologically inspired value-driven learning into autonomous robots to simulate foraging behavior (Almassy, Edelman, & Sporns, 1998; Sporns & Alexander, 2002, 2003). One prominent recent theory of reward-learning is that the release of dopamine from midbrain areas correlates with the onset of unexpected reward (Schultz, 2000). The baseline sensitivity of this dopamine release mechanism to potentially rewarding stimulus cues may play a factor in the degree to which an organism engages in reward-seeking and dominant behavior with increased sensitivity correlating with increasing predisposition towards “extraverted” behavior (Depue & Collins, 1999). Connectionist AI and computational neuroscience researchers have already provided many useful insights into the utility of supervised and unsupervised learning and are continuing to develop more biologically plausible algorithms for their models (O'Reilly & Munakata, 2000). The principle suggested by Level 2.0 is that: *Learning of new behaviors through reinforcement is an important component of human-level volition*. The advent of general purpose robots capable of learning complex behaviors through a reinforcement training in the manner of animal development (Weng et al., 2001) will be a major technological advance.

Working memory is the main architectural feature that needs to be developed for Level 2.1 (Ideational Organisms). (O'Reilly & Frank, 2006) provides a recent example of a biologically inspired neurocomputational model of working memory. This model, which uses an actor-critic architecture, suggests a way in which prefrontal cortical and basal ganglia areas of mammalian brain may learn how to mediate working memory tasks. Such a model is currently being used as the basis for adding working memory to robots (Skubic, Noelle, Wilkes, Kawamura, & Keller, 2004). The focus of the author’s own research (Chadderdon & Sporns, 2006) has involved developing a neurocomputational model of working

memory and its relationship to task-oriented behavior selection. Although this work has developed a model which is capable of executing both delayed match-to-sample (DMS) and delayed non-match-to-sample (DNMS) tasks using working memory to remember the task and target stimulus, there is yet no learning component of the model, so it is effectively a Level 1.1 system. The author is currently seeking to integrate the current model with a learning mechanism so that the model may learn the DMS and DNMS tasks from scratch based on reinforcement. This will move the system significantly in the direction of Level 2.1. Studies of prefrontal cortex (Miller & Cohen, 2001) and basal ganglia (Gurney, Prescott, & Redgrave, 1998, 2001) are likely to be important in the development of biologically inspired Level 2.1 mechanisms. Level 2.1 demonstrates the principle that: *Working memory is an important component of human-level volition.* With working memory and reinforcement learning, an organism becomes able to learn tasks, meaning that it may learn to temporarily shift its mode of behavior to meet a goal that is currently relevant at the expense of other behaviors that might ordinarily be prompted reflexively by current stimuli. Thus, it becomes no longer stimulus-driven. Robots that are capable of Level 2.1 functionality will be highly versatile, able to learn not only stimulus-response patterns, but tasks in the sense given above.

For Level 2.2 (Recollective Organisms), the main architectural feature that needs to be developed is long-term (i.e. episodic and semantic) memory to allow, for example, case-based action selection. (Verschure et al., 2003) gives an example of how such case-based reasoning might work to control an autonomous robot. Their distributed adaptive control (DAC) architecture has a contextual control layer which includes both short-term and long-term memory components that become active when the system has learned sufficiently consistent mappings between conditioned stimuli (CSs) and conditioned responses (CRs). The STM component chronicles at each time step the current combination of CS and resulting CR, and when a goal is reached, the current sequence in STM is copied into LTM. When the DAC system isn't reacting to unconditional stimuli and the contextual control layer is active, an associative pattern match is done between the current CS and the CS in the CS/CR pairs in LTM. The closest match triggers the corresponding CR behavior allowing case-based response of the system.

Choice from within a particular LTM sequence biases selection for action from later links in that sequence. Such an LTM architecture that stores episodic events that may be cross-linked into chains may have a correlate in the functioning of the hippocampus (Eichenbaum, 2000). (Burgess, Donnett, Jeffery, & O'Keefe, 1997) provides an example of a robotic model of how hippocampal place cells may form and how they may guide navigation in an open environment. In general, study of the hippocampus and its mechanisms is likely to be of importance in developing Level 2.2 organisms. The principle suggested here is that: *Long-term memory is an important component of human-level volition.* Effective robot navigation may well depend on Level 2.2 functionality, and certainly, if the robot is to have a sense of experiential history, it will need to have a functioning long-term memory.

The functionality of Level 2.3 (Deliberative Organisms) is, at least in part, a matter of coordinating working memory and long-term memory in such a way that deliberation over different candidate hypotheses or behaviors is allowed. Ironically, a number of classical AI inspired systems already perform a kind of deliberation because classical AI is based mainly on search of a state space for a viable goal state. A chess program that does a tree search in an attempt to find its best move according to a heuristic fitness function is iteratively trying and evaluating candidate hypotheses in Level 2.3 fashion. However, neither the heuristic function, nor the repertoire of behaviors of the system are learned, nor generally is the heuristic function modified by the results of its deliberation. The chess-playing system cannot learn other behaviors outside of the narrow domain of chess, either. Whatever volition Deep Blue possesses is limited only to chess-playing, and that is hard-coded by the programmer and not open to self-modification.

The author, at time of writing, did not find any obvious studies of the neurocomputational mechanisms of deliberation, though the models of working memory and hippocampus previously mentioned could certainly constitute parts of this mechanism. Sloman (2003) points out the importance in deliberation of the interaction of STM (or working memory) and LTM, so prefrontal cortex, basal ganglia, hippocampus, as well as other brain regions such as anterior cingulate cortex are likely to mediate deliberation, though investigations are currently at an early stage. The final principle to be

suggested here is that: *Deliberation among alternative possibilities is essential for human-level volition*. A robot which is capable of Level 2.3 functionality will be able to learn tasks, keep an episodic record of its experience, and use its previous experiences to pre-test candidate behaviors offline before committing to an action. When such robots exist and are capable of a sufficient array of behaviors, they will probably give the observer an almost palpable sense of being alive, even lacking natural language capabilities.

In our view, the current focus of AI should be on the various sublevels of Level 2 (Contained Self). Level 2.3 (Deliberative Organisms) is a quintessential level of volition such as might be seen in mammals, perhaps the foundational level that will need to be achieved before systems may be developed that consistently pass the Turing test. Development of the social and manipulative scaffolding that will be needed for Level 3.0 and 3.1 organisms, respectively, may be attempted early, but many of the interesting behaviors we'd expect to see at these levels will be absent if the systems lack deliberative capabilities as individuals. In addition, it may be doubtful that true symbolic usage such as is seen at Level 3.2 (Symbolic Organisms) will even be a possibility unless the organism has first achieved the previous volitional levels. This is because deliberation and social scaffolding are critically important in symbolic communication, and manipulation of objects may provide much motivation for the action/verb-components of language. Level 3.3 (Cultural Organisms), of course, builds immediately on Level 3.2.

Clearly, there remains much to accomplish in AI before android volition (and beyond) is achieved. However, in the long run there seems plenty of cause to be optimistic. Computer technology continues to improve both in processing speed and information-holding capacity. Inquiries in the neurosciences daily reveal more about the mechanisms of human and animal behavior, and computational modelers continually strive to integrate this information, fill in gaps of understanding (and data), and infer possible mechanisms. A relatively new discipline, cognitive science, has emerged which is attracting the attention of many different disciplines and fostering a great deal of cross-disciplinary communication. This concerted effort to reveal the workings of human and animal minds will pay dividends for those who are trying to engineer intelligence and behavior in computational systems.

Endnotes

1. A more detailed explanation of the functionalist and other positions regarding the solution of the mind/body problem can be found in (Churchland, 1996).

2. In Dennett's view, one of the major errors made by philosophers regarding free will is to define "could have been otherwise" as meaning that, given *exactly* the same previous state of the universe, a different result might have happened (Dennett, 1984, 2003). Doing this has a tendency to define freedom of the will out of existence if you believe the universe is deterministic.

3. Note, similar states of the *environment*, not the universe as a whole. If the environment and the mental states of the organism were both similar, we might likely see the same resulting behavior (assuming small chaotic sensitivity to the initial difference). The key is that the internal state drives which behavior gets enacted, not the environment alone.

4. 'Adaptive' also has the connotation of involving learning, but here we intend only to use it in the sense of meeting a goal successfully. The organism is adapting to the environment, in a small sense, through its present behavior, but not necessarily remembering the act or learning from it, which would be adaptation in a larger, higher-order, sense.

Acknowledgements

G.C. was supported by NIMH training grant T32MH019879-11 and by an IU Cognitive Science fellowship. Comments by Will Alexander, Andy Clark, Jerry Guern, Chris Honey, Gary Lucas, Olaf Sporns, Karola Stotz, Julie Stout, Peter Todd, and Jim Townsend on earlier drafts of this paper are gratefully acknowledged.

References

Webster's New World Dictionary & Thesaurus. (1998). (2.0 ed.): Accent Software International. Macmillan Publishers.

Honda Worldwide | ASIMO. (2008). Retrieved January 12, 2008, from http://world.honda.com/ASIMO/

Jabberwacky. (2008). Retrieved January 12, 2008, from http://www.jabberwacky.com/

A.L.I.C.E. Artificial Intelligence Foundation. (2008). Retrieved January 12, 2008, from http://www.alicebot.org/

Almassy, N., Edelman, G. M., & Sporns, O. (1998). Behavioral constraints in the development of neuronal properties: a cortical model embedded in a real-world device. *Cerebral Cortex*, *8* (4), 346-361.

Anderson, M. L. (2005). How to study the mind: an introduction to embodied cognition. In F. Santoianni & C. Sabatano (Eds.), *Embodied Cognition and Perceptual Learning in Adaptive Development*.Newcastle upon Tyne: Cambridge Scholars Press.

Bäck, T. (1996). *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms*.New York: Oxford University Press.

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: a cross-cultural study. *Evolution and Human Behavior, 26*, 313-331.

Boden, M. A. (1996). Autonomy and artificiality. In M. A. Boden (Ed.), *The Philosophy of Artificial Life* (pp. 95-108). Oxford: Oxford University Press.

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*.Cambridge, MA: MIT Press.

Brooks, R. A. (1991). Intelligence without representation. In J. Haugeland (Ed.), *Mind Design II* (pp. 395-420). Cambridge, MA: MIT Press.

Burgess, N., Donnett, J. G., Jeffery, K. J., & O'Keefe, J. (1997). Robotic and neuronal simulation of the hippocampus and rat navigation. *Philosophical Transactions of the Royal Society of London B Biological Science, 352*, 1535-1543.

Chadderdon, G. L., & Sporns, O. (2006). A large-scale neurocomputational model of task-oriented behavior selection and working memory in prefrontal cortex. *Journal of Cognitive Neuroscience, 18*(2), 242-257.

Chisholm, R. (1964). Human freedom and the self. In G. Watson (Ed.), *Free Will* (pp. 24-35). Oxford: Oxford University Press.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. In R. Cummins & D. D. Cummins (Eds.), *Minds, Brains, and Computers* (pp. 500-512): Blackwell Publishing.

Churchland, P. M. (1996). *Matter and Consciousness*.Cambridge, MA: MIT Press.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*.Cambridge, MA: MIT Press.

Clark, A. (2001). *Mindware: An Introduction to the Philosophy of Cognitive Science*.New York, NY: Oxford University Press.

Clayton, N. S., Emery, N. J., & Dickinson, A. (2006). The prospective of food caching and recovery by western scrub-jays. *Comparative Cognition and Behavior Reviews, 1*, 1-11.

Deacon, T. W. (1997). *The Symbolic Species: the co-evolution of language and the brain*.New York, NY: W. W. Norton and Company.

Dennett, D. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*.Cambridge, MA: MIT Press.

Dennett, D. (1991). Real patterns. *The Journal of Philosophy, 88*(1), 27-51.

Dennett, D. (1996). *Kinds of Minds*.New York, NY: Basic Books.

Dennett, D. (2003). *Freedom Evolves*.New York, NY: Viking.

Depue, R. A., & Collins, P. F. (1999). Neurobiology of the structure of personality: dopamine, facilitation of incentive motivation, and extraversion. *Behavioral and Brain Sciences, 22*(3), 491-517; discussion 518-469.

Dreyfus, H. L. (1979). From micro-worlds to knowledge representation: AI at an impasse. In J. Haugeland (Ed.), *Mind Design II* (pp. 143-182). Cambridge, MA: MIT Press.

Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nature Reviews Neuroscience, 1*, 41-50.

Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*.Cambridge, MA: MIT Press.

Fuller, E. (1959). *Bulfinch's Mythology (abridged)*.New York, NY: Dell Publishing.

Gurney, K. N., Prescott, T. J., & Redgrave, P. (1998). *The basal ganglia viewed as an action selection device.* Paper presented at the 8th International Conference on Artificial Neural Networks, Skövde, Sweden.

Gurney, K. N., Prescott, T. J., & Redgrave, P. (2001). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics, 84*, 401-410.

Han, C. J., O'Tuathaigh, C. M., van Trigt, L., Quinn, J. J., Fanselow, M. S., Mongeau, R., et al. (2003). Trace but not delay fear conditioning requires attention and the anterior cingulate cortex. *Proceedings of the National Academy of Sciences, 100*(22), 13087-13092.

Idel, M. (1990). *Golem: Jewish Magical and Mystical Traditions on the Artificial Anthropoid*.Albany, NY: State University of New York Press.

Kane, R. (1996). *The Significance of Free Will*.Oxford: Oxford University Press.

Langton, C. G. (1989). Artificial life. In C. G. Langton (Ed.), *Artificial Life I: Proceedings of the Workshop on Artificial Life, Los Alamos, New Mexico, September 1987* (pp. 1-47). Reading, MA: Addison-Wesley.

Levy, S. (1992). *Artificial Life: A Report from the Frontier Where Computers Meet Biology*.New York, NY: Vintage Books.

Löffler, K., Gienger, M., & Pfeiffer, F. (2003). *Sensor and control design of a dynamically stable biped robot.* Paper presented at the International Conference on Robotics and Automation.

MacLean, P. D. (1990). *The Triune Brain in Evolution: Role in Paleocerebral Functions*.New York, NY: Plenum Press.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review in Neuroscience, 24*, 167-202.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation, 18*, 283-328.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience*.Cambridge, MA: MIT Press.

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotion*.New York, NY: Oxford University Press.

Pepperberg, I. M. (2002). Cognitive and communicative abilities of grey parrots. *Current Directions in Psychological Science, 11*(3), 83-87.

Pfeifer, R., & Scheier, C. (1999). *Understanding Intelligence*.Cambridge, MA: MIT Press.

Rich, E., & Knight, K. (1991). *Artificial Intelligence* (Second ed.): McGraw-Hill, Inc.

Russell, S. J., & Norvig, P. (2002). *Artificial Intelligence: A Modern Approach* (Second ed.). Upper Saddle River, NJ: Prentice Hall.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences, 4*(8), 299-309.

Schultz, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience, 1*, 199-207.

Shelley, M. (1818). *Frankenstein: Or, the Modern Prometheus*.New York, NY: Penguin Books.

Skubic, M., Noelle, D., Wilkes, M., Kawamura, K., & Keller, J. (2004). *Biologically inspired adaptive working memory for robots.* Paper presented at the AAAI 2004 Fall Symposium, Workshop on The Intersection of Cognitive Science and Robotics: From Interfaces to Intelligence, Washington, D.C.

Sloman, A. (2003). How many separately evolved emotional beasties live within us? In R. Trappl, P. Petta & S. Payr (Eds.), *Emotions in Humans and Artifacts* (pp. 35-114). Cambridge, MA: MIT Press.

Sony Corporation. (1999). Sony Launches Four-Legged Entertainment Robot. Retrieved January 12, 2008, from http://www.sony.net/SonyInfo/News/Press_Archive/199905/99-046/index.html

Sporns, O., & Alexander, W. H. (2002). Neuromodulation and plasticity in an autonomous robot. *Neural Networks, 15*, 761-774.

Sporns, O., & Alexander, W. H. (2003). Neuromodulation in a learning robot: interactions between neural plasticity and behavior. *Proceedings from the International Joint Conference on Neural Networks 2003*, 2789-2794.

Turing, A. (1950). Computing machinery and intelligence. In J. Haugeland (Ed.), *Mind Design II* (pp. 29-56). Cambridge, MA: MIT Press.

Verschure, P. F. M. J., Voegtlin, T., & Douglas, R. J. (2003). Environmentally mediated synergy between perception and behavior in mobile robots. *Nature, 425*, 620-624.

Weng, J., McClelland, J. L., Pentland, A., Sporns, O., Stockman, I., Sur, M., et al. (2001). Autonomous mental development by robots and animals. *Science, 291*, 599-600.

Yaeger, L. S. (1993). Computational genetics, physiology, metabolism, neural systems, learning, vision, and behavior. In C. G. Langton (Ed.), *Artificial Life III: Proceedings of the Workshop on Artificial Life, Santa Fe, New Mexico, June 1992* (pp. 263-298). Reading, MA: Addison-Wesley.

Figure Captions

Figure 1. Overview of volitional scale for measuring artificial and natural systems. The levels, shown on the left, and their sublevels, shown in the middle, build features on top of the lower (sub-)levels. Non-volitional systems (further described in Table 1) are characterized by their lack of tropistic behavior. Instinct-driven automata (see Table 2), by contrast, are characterized by 'hard-wired', 'instinctive' tropistic behavior patterns which can be considered, in a rudimentary fashion, value-/goal-driven. Contained self organisms (see Table 3), in addition to instinctive tropisms, have the capacity for learning new behaviors from experience, but have limited capabilities for environmental influence which means that their self, their agency, is more self-contained. Finally, extended self organisms (see Table 4) have an influence on and an interactivity with their environments that makes it reasonable to talk about their self as being (at least partially) extended into their environments (as, for example, a couple or group of individuals working closely together, or a person using a note-pad to improve their mnemonic capacity). The text to the right characterizes how each sublevel differs from its preceding sublevel. Each sublevel has a "litmus question" (see Tables and Section 3) which allows categorization of the systems whose volitional endowment we want to measure.

Table Captions

Table 1. Features of Level 0: Non-Volitional Systems.

Table 2. Features of Level 1: Instinct-Driven Automata.

Table 3. Features of Level 2: Contained Self Organisms.

Table 4. Features of Level 3: Extended Self Organisms.

Biography

George L. Chadderdon holds a B.S. in Computer Engineering (University of Illinois, Urbana-Champaign) and an M.S. in Computer Science (University of California, San Diego). He is currently completing a doctorate in Psychology and Cognitive Science at Indiana University in Bloomington as a student of Olaf Sporns. His academic interests include artificial intelligence, computational neuroscience, and the mechanisms that underlie individual differences in temperament and personality; and he has recently been studying the dopaminergic mechanisms involved in working memory and reinforcement learning. His outside interests include creative writing, guitar, and musical composition.

Figure 1. Overview of volitional scale

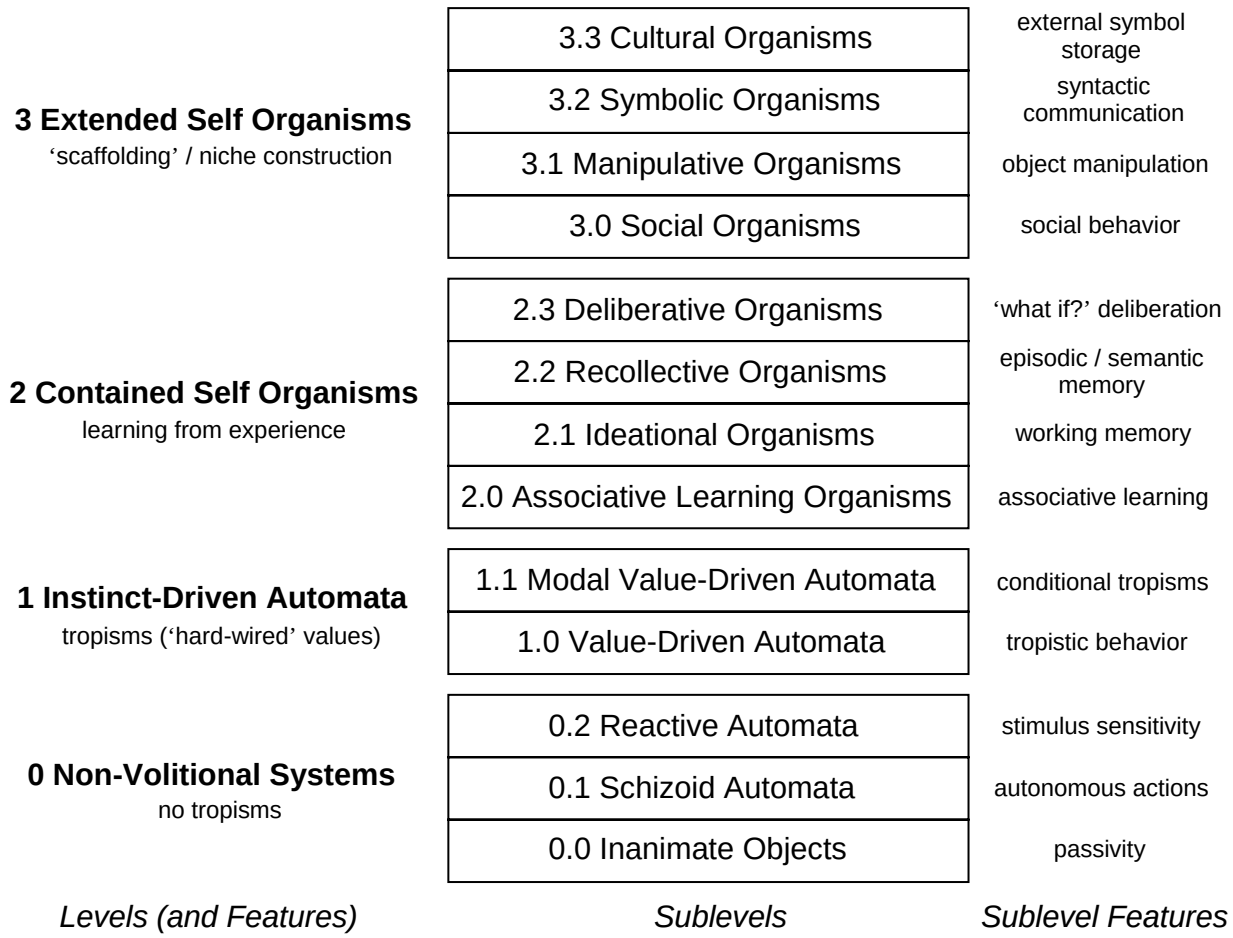

Table 1. Level 0: Non-Volitional Systems

Level 0.0	**Inanimate Objects**
Critical Attribute/s	N/A
Litmus Test	This is the default level for any object or system.
Exemplar Systems	rocks, cups, spoons, particles

Level 0.1	**Schizoid Automata**
Critical Attribute/s	continuous autonomous behavior (actions)
Litmus Test	Does the system move or act on its own, i.e., without continuous prompting by external forces?
Exemplar Systems	clocks, wind-up dolls

Level 0.2	**Reactive Automata**
Critical Attribute/s	conditional autonomous behavior (stimulus input)
Litmus Test	Is the system's spontaneous behavior modified by events/conditions in the environment?
Exemplar Systems	vehicle engines, motors (when running)

Table 2. Level 1: Instinct-Driven Automata

Level 1.0	**Value-Driven Automata**
Critical Attribute/s	difference minimizing/maximizing (i.e. goal/tropistic) behavior
Litmus Test	Does the system appear to be trying to approach or avoid any object or occurrence of an event in its environment?
Exemplar Systems	thermostats, heat-seeking missiles

Level 1.1	**Modal Value-Driven Automata**
Critical Attribute/s	behavioral modes (driven by environment or internal state)
Litmus Test	Does the system have different sets of goals active during different environmental or bodily conditions?
Exemplar Systems	single-celled organisms, sea-slugs, Pac-man ghosts

Table 3. Level 2: Contained Self Organisms

Level 2.0	**Associative Learning Organisms**
Critical Attribute/s	associative learning (reinforcement, conditioning, etc.)
Litmus Test	Does the system develop new adaptive approach or avoidance patterns over time?
Exemplar Systems	simple reactive animals that can learn preferences (e.g. some insects)

Level 2.1	**Ideational Organisms**
Critical Attribute/s	dynamic short-term memory, ideation
Litmus Test	Can the system engage in a task that requires working memory (e.g. delayed non-match-to-sample)?
Exemplar Systems	animals that can hold items in memory for task behaviors

Level 2.2	**Recollective Organisms**
Critical Attribute/s	long-term (episodic and semantic) memory
Litmus Test	Can the system engage in a task that requires long-term memory?
Exemplar Systems	animals that remember semantic relationships or gestalt events

Level 2.3	**Deliberative Organisms**
Critical Attribute/s	‘what if’ decision-making, attentional control of memory access
Litmus Test	Can the system engage in a behavior (e.g. game-playing, navigation) that requires evaluation of multiple possibilities without action?
Exemplar Systems	animals that can navigate complex spaces

Table 4. Level 3: Extended Self Organisms

Level 3.0	**Social Organisms**
Critical Attribute/s	social signaling and representation
Litmus Test	Does the organism send and selectively respond to social cues?
Exemplar Systems	rats, dogs, cats, horses, etc.

Level 3.1	**Manipulative Organisms**
Critical Attribute/s	object manipulation
Litmus Test	Can the system pick up and move around objects in its environment?
Exemplar Systems	monkeys and apes

Level 3.2	**Symbolic Organisms**
Critical Attribute/s	symbol/language usage
Litmus Test	Does the system communicate using language that has syntax as well as semantics?
Exemplar Systems	primitive humans

Level 3.3	**Cultural Organisms**
Critical Attribute/s	external symbol storage (e.g. writing)
Litmus Test	Does the system use external storage of symbols as a repository for knowledge?
Exemplar Systems	modern humans