Supplementary Material for:

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

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

S1 Detailed Description of Volitional Scale Sublevels

S1.1 Level 0: Non-Volitional Systems

S1.1.1 Level 0.0: Inanimate Objects

Objects in this level can be said to have no volition whatsoever because they enact no behavior that could be called autonomous. They respond passively, though continuously, to forces in the environment, but do nothing not immediately dependent on those external forces. Examples include rocks, footballs, ball-point pens, and individual particles.

S1.1.2 Level 0.1: Schizoid Automata

When a system gains the capacity for autonomous behavior, it reaches this level. Schizoid automata do not engage in goal-oriented behavior, but can spontaneously update an internal state or engage in sustained continuous behavior. Behavior may be rhythmic or even random. Examples include clocks and wind-up dolls since these act spontaneously, but in a manner generally independent of any stimuli. Properly speaking, it might be argued that such systems do not possess volition because they do not discriminate different states of the universe and have not even innate goals. Because of their suggestively spontaneous, internally-driven behavior, however, we will say that such systems possess *proto-volition.*

S1.1.3 Level 0.2: Reactive Automata

The next innovation at Level 0 is to have the environment influence the spontaneous behavior of the system in some way. Reactive automata can be said to actively engage in behaviors, but these behaviors can now be modified by environmental cues. Imagine, for example, the engine of an automobile while it is running. It will generate rotary motion quite spontaneously, but will respond to opening and closing of the fuel intake so as to change its speed according to how far the accelerator pedal is pressed. Another example might be a robot that moves in a straight line, but executes a random turn when you clap your hands. The sphere of proto-volition has expanded to include sensitivity to the environment. It now makes sense to talk about the system having stimuli it responds/reacts to.

S1.2 Level 1: Instinct-Driven Automata

S1.2.1 Level 1.0: Value-Driven Automata

Here we begin to see a kind of phenomenon that we may cease calling mere proto-volition. These objects have what might be called (implicit) goals, or value-oriented behavior. To have a goal/value implies having some target to be approached or avoided. The system is trying to either minimize (approach) or maximize (avoid) some difference between a current state and target state in the environment. (That the goal state is external to the agent is important.) Thermostats and heat-seeking missiles exist at this level. Braitenberg (1984) discusses simple robots that are wired to chase after or avoid light sources. An interesting property of these robots is that there is no explicit symbolic representation of a goal contained in them—light-sensors are wired directly to the left and right motors driving the wheels. Nonetheless, these would be considered to possess volition at the level defined here. Volition is expanded because we can now talk about the system having goals, desires, intentions, etc., though such simple systems probably cannot be said to be aware of their goals as are higher vertebrates.

S1.2.2 Level 1.1: Modal Value-Driven Automata

The next volitional innovation to add is the ability of the environment and/or internal state of the system (e.g. hunger) to affect which goals are active at any given time. Stimuli and internal states select, in effect, what *mode* the organism is in, and the particular goal or set of goals that is active will depend just on this mode setting. Other than that, the system has no memory or learning. Effectively, these systems must be designed either by hard-wiring or evolutionary processes since no learning is possible. Single-celled organisms and (to an approximation) insects seem to fall into this level (though at least some insects, for example honeybees and ants, may have some reinforcement learning capabilities (Cammaerts, 2004; Papaj & Lewis, 1992; Wersing, Grünewald, & Menzel, 2003)). Video game “bots”, too, often operate at this level. The ghosts in Pac-man are an illustrative example of systems at this volitional level. The four ghosts each have two modes, essentially: EatPlayer or RunFromPlayer. In the first, default state a ghost will move towards the player (approach) until it eats them. When the player

eats a power-up pellet, however, the situation has changed and the ghost runs away from the player (avoidance) in order to avoid being eaten by the player.

Volition in Level 1.1 has been expanded by goals, desires, and intentions being appropriately (if reflexively) selected by current conditions, either in the environment or in the system's internal state. Interestingly, such a system may be programmed to act very differently depending in a hard to directly observe way on both environmental and internal (i.e. bodily) context. Seeing the system act differently in outwardly similar situations, an observer (adopting the intentional stance) might well say that the organism made a different *choice* under the different occurrences. The Pac-man ghost at any given moment has the implicit choice of chasing or running away from the player. Which it does will depend on whether its behavior mode is in a certain state, i.e., which represents whether a power-pellet is in effect and the ghost is therefore vulnerable.

S1.3 Level 2: Contained Self Organisms

S1.3.1 Level 2.0: Associative Learning Organisms

The boundary between this and earlier levels could probably be considered the Great Divide of volition, and it corresponds with the boundary between the first (*Darwinian creatures*) and second level (*Skinnerian creatures*) of Dennett's Tower of Generate-and-Test scale (Dennett, 1996). Learning, especially reinforcement learning, is the next innovation for our volitional systems. The ability to learn stimulus features and associations accompanies reinforcement learning and probably utilizes the same mechanisms. Classical AI has concerned itself too infrequently with learning, and the learning of autonomous behaviors even more infrequently. Expert systems such as Mycin (Rich & Knight, 1991) and Deep Blue (McDermott, 1997) fail to convince us that they possess volition, not only because they do not locomote or act spontaneously, but because they engage in no learning. (On the other hand, many learning systems that exist, including many connectionist systems, have learning, but no autonomous behavior.) Video-game bots like the Pac-man ghosts or like the opponents in "first-person shooter" games like Quake or Unreal act with life-like spontaneity and cunning purpose, but their behavior is mostly reactive and devoid of learning. Most vertebrate animals, however, exist at least at this level of

volition and possess the capacity for learning. (Indeed, most invertebrates probably also possess some reinforcement learning capacity because of the inherent plasticity of neural systems.)

The sphere of volition has increased for these systems because they may now adapt to their environments, learning goals and strategies during their lifetime rather than being limited by their hard-wiring. The ability to modify one's goals and strategies rather than always acting the same way every time the same stimulus is encountered is a crucial component of higher volition. Level 1.1 systems may be hard-wired to engage in fit choices in response to their environments, but Level 2.0 systems may in fact learn what particular choices are good to make (or not) based on experience. They may actively explore their environments (perhaps through hard-wired exploratory behavior patterns) and learn to make "wiser" (i.e., more adaptive) choices, by learning what stimuli and behaviors are helpful or harmful.

Organisms of the same species, even if initially wired identically, may develop their own "personalities" by acquiring idiosyncratic likes and dislikes through differences in the reinforcement patterns in their respective environments. Thus, individuality (beyond genetic variation) is born at this level. Just as reinforcement makes learning of representation of stimulus features possible, it would allow the representation of individual conspecifics to be made possible, meaning Level 2.0 organisms could develop a representation of the identity of other organisms of the same kind. Such organisms could learn to prefer or shun the company of other individuals based on how their presence and behavior affects them. All of this could be possible without any self-awareness or self-representation whatsoever.

Another feature added by learning is the potential for predictive behavior. Through such mechanisms as temporal difference (TD) learning (Suri, 2002; Sutton & Barto, 1998), a system might learn whole sequences of behavior necessary to accomplish some goal, and also the ability to predict what kinds of stimuli precede rewards and punishments. Early detection of impending reward or punishment could then activate approach or avoidance behaviors, respectively.

S1.3.2 Level 2.1: Ideational Organisms

The next volitional innovation for organisms is possession of a reusable short-term memory which may be used to buffer information which might make a difference in the execution of behaviors.

This kind of short-term memory is known as *working memory* (Baddeley, 2003) and can be thought of as a kind of "blackboard" memory the organism can use to store information behaviors might require during execution. With working memory comes the ability for an organism to fixate on some perceivable stimulus or environmental condition cue and hold a representation of it in memory when it disappears.

The ability of working memory to represent, or *ideate*, previously-evidenced stimuli or conditions in the environment allows a number of cognitive and control features to Level 2.1 organisms unavailable to those at Level 2.0. Distinct goals/tasks may be ideated, as when a transient cue in the environment (e.g. a blue light) signals the need of a rat to approach its food-dispenser whereas another cue (e.g. a tone or a red light) signals the need of the rat to run from the food-dispenser part of its cage to avoid an electric shock. Subgoals may be maintained also, such as, perhaps, the need of a chimp to have a wooden box located under a banana tree before it can climb on it to reach the fruit. When the subgoal is completed, its representation may be released and the chimp will proceed to the next foraging stage. Not only tasks and subtasks may be cued, but also targets for such tasks, such as when an animal is shown a particular object it is supposed to fetch and it is able to retrieve the object from among a number of other irrelevant objects. The *manner* in which a task is executed may also be ideated, as when a particular object to be fetched may require a different grip or search strategy depending on other transiently-cued conditions in the environment. Finally, ideation may be used to bias the perceptual and attentional processes of the organism such that they are more likely to perceive certain stimuli rather than others during the period of ideation.

Organisms at this level still lack long-term memory, so they cannot really engage in deliberation and hypothetical 'what if' reasoning (Sloman, 2003). However, the ability to fix representations in mind allows a new flexibility because the organism is no longer exclusively driven by the immediate stimulus in the environment in conjunction with the previous reinforcement patterns. Now, the organism can act on cues that have disappeared but still have lingering importance. Its reinforcement learning may allow it to learn adaptive ideational states that are selected by particular transient environmental cues.

S1.3.3 Level 2.2: Recollective Organisms

The next volition-increasing innovation is to provide the organism with a functioning long-term *declarative* memory, i.e., both *semantic* (static facts and conceptual relations) and *episodic* (experiences and dynamic events) memory. Without this, an organism's volition and consciousness are trapped in the present. For organisms that exist in environments that afford goal-satisfaction by random walk, declarative memory probably isn't needed, but most environments do not have an even distribution of resources, so declarative, particularly episodic, memory becomes a useful way of encoding representations of place. Episodic memory can be imagined crudely as a kind of independent chronicler that's always running the background remembering specific events, places, etc., along with valuational states associated with these things. When the organism encounters a situation that matches closely enough a remembered episode, the chronicler may not only signal familiarity, but may also fill in information missing from the current stimuli with information from the matching episode. This filled-in information might include a prediction of the sort of emotional state that will soon happen or it might include specific motor plans (or working memory ideational representations) that were used in the past when the situation was encountered. Recollective organisms would thus gain the ability to occasionally draw on remote past experiences to make simple decisions, a kind of case-based decision-making where the closest-matching case sets the response.

S1.3.4 Level 2.3: Deliberative Organisms

The next step for organisms is the ability to engage in deliberative 'what if' reasoning. Sloman (2003) regards this faculty as being a layer of processing which modulates the activity in a reactive processing layer. Instead of merely picking the behavior that corresponds to the resolution in the organism's closest-matching previous experience, deliberation involves "trying out" different action possibilities without actually performing them, and is a very adaptive ability to have when trial and error may result in death or injury. Thus with this level, we arrive at Dennett's *Popperian creatures* Tower of Generate-and-Test level (Dennett, 1996).

Merely possessing short- and long-term memory is not enough to allow deliberation. First, there needs to be some mechanism of disengaging the organism's mental processes (at least partially) from both external stimuli and from motor activity. Otherwise, the act of imagining a particular course of action might be disrupted by current stimuli or might trigger the action itself with potentially disastrous consequences. Second, in order to strategically direct access to episodic memory, there needs to be some kind of executive processing that focuses attention on remembered episodes pertinent to the current task or goal.

When the organism needs to decide on a particular course of action, it might generate a candidate action (or several) in working memory. These candidates could be passed in sequence into episodic memory as incomplete patterns to be filled in. Similar previously occurring actions/events in the organism's life might cause the patterns to be completed with the appropriate emotional valence representations, and downstream processes could decide from this whether the candidate action is a good or bad idea. Information coming from episodic memory might also pattern complete in a way which specifies next-step actions, as in the case of navigation in a complex environment in order to reach the known location of a food source.

In fact, deliberative abilities may have evolved in order to allow intelligent navigation through complex and potentially hostile spaces. Episodic memory (through such a mechanism as hippocampal "place cells") might store associations between particular places, hazards (e.g. quicksand) or resources (e.g. food), and appropriate motor behaviors performed in the places (Clark, 1997; Mataric, 1991). A hungry rat's episodic memory might complete the goal pattern of its hunger (G_0) in such a way that it includes the location of one food source in the environment. This location might be stored in working memory as a subgoal (G_1). G_1 may then be sent off to episodic memory, and the associated locations leading immediately to G_1's location may be found and stored in working memory. The best of these, as determined by iterative emotional evaluation by submission to episodic memory, would be saved as subgoal G_2. Using this iterative, backward-chaining process, the organism could assemble an appropriate path from where it sits to the food source.[1] Following this path would then amount to executing the

sequence of subgoals G_n, G_{n-1}, etc., and releasing their representations, until the food is in sight (G_0). Thus, a deliberative organism would be capable of the kind of goal-stack planning we often engage in.

In theory, it seems possible that organisms that have achieved this level of volition may be able to deliberate on previous events in their lives and use emotional valences of these events to modify their tendencies to engage in particular actions in similar contexts in the future (the benefit of remorse or reminiscing on triumphs). The ability to reflect on one's past and make adjustments in one's future strategies as a result is a quintessentially human feature and a hallmark of advanced volitional capacity.

S1.4 Level 3: Extended Self Organisms

S1.4.1 Level 3.0: Social Organisms

The key innovation at this level is extensive social signaling and representation. Rats, dogs, horses, etc. all seem to exhibit these features. Social signaling and interpretation, in combination with reinforcement learning, allows organisms to learn representations of their social as well as physical environment, and allows organisms to learn behaviors that influence other organisms in desired ways. In short, it allows influencing behaviors or cooperation to be learned. Extensive social behavior not only provides benefits through (relative) safety in numbers, but also can potentially allow individuals within the group to specialize in particular roles that improve the chances of survival for most or all individuals in the group. Deliberative 'what if' thinking allows animals to plan out interactions with conspecifics and maybe other species as well. For example, if the organism faces a challenge, it may deliberate on the likelihood of various herd-mates being willing and able to aid them, and may decide which one to solicit before approaching them. The sphere of volition expands because recruiting cooperation from conspecifics is a strategy that can greatly improve chances of meeting goals. It also extends the range of influence of the organism since organisms in a social network may be widely spread spatially.

S1.4.2 Level 3.1: Manipulative Organisms

The next major volitional innovation, and one possessed most clearly by primates, is the ability to physically manipulate objects in their environment. This allows the organism to arrange resources in a way which is convenient for them and also to use tools. Social signaling and learning in combination

with this allows organisms to learn manipulative behaviors from each other, including tool-making. Chimpanzees and other great apes are the archetypes of this level (Matsuzawa, 2002). Although they lack language, they can be said to have a simple kind of tribal structure where each of the members have their own distinct individual personalities, and each tribe might have its own particular customs: for example, methods for washing wheat or potatoes, cracking nuts, or appropriate grooming behaviors (Dautenhahn, 2000; Matsuzawa, 2002). The sphere of volition is expanded by allowing systems to make more thoroughgoing changes to the environment in order to meet goals. Changes made to the environment may also be shared among conspecifics, augmenting the adaptive effects of socialization.

S1.4.3 Level 3.2: Symbolic Organisms

Finally, we arrive at the lowest level occupied by humans. The key innovation at this level is symbol use, in particular, spoken language. Here we may recognize primitive man: man prior to the invention of writing. Spoken language allows both better communication with others of the same species and also another means of self-representing. With language comes the ability to create and retain in memory instructive or fanciful narratives which may be passed on by word of mouth. Art and religion probably begins at this level as well as ancient verse literature, including epic and song. Sophisticated cultures can be developed by organisms at this level because customs can be taught verbally rather than acquired by motor imitation. The sphere of volition expands through better communication of information, both to conspecifics and within oneself through self-talk.

S1.4.4 Level 3.3: Cultural Organisms

At this point, we arrive at our current level of volition (at least as individuals). External symbol storage is the next innovation, i.e., written language. With written language, it is possible to codify transactions, laws, customs, narratives, technological methods, and many other important fixtures of civilization. Systematic philosophy, science, and literature are now possible, and knowledge can be carried over great distances and many generations. The sphere of volition expands greatly over time and space.

S2 Examples of Measurement Using the Scale

S2.1 Example 1: A Pac-man Ghost

For our first example, we revisit the case in Section S1.2.2 of a ghost computer adversary in Pac-man. First, we assume it's inanimate, i.e., at Level 0.0. Then we proceed to Level 0.1 and ask: "Does the system move or act on its own, i.e., without continuous prompting by external forces?" The answer is clearly Yes since all of the ghosts can be observed moving on the screen, so a ghost is at least a schizoid automaton. So we proceed to Level 0.2 and ask: "Is the system's spontaneous behavior modified by events/conditions in the environment?" As we move Pac-man around, the ghosts follow him, so the answer is Yes. Therefore, we proceed to Level 1.0 and ask: "Does the system appear to be trying to approach or avoid any object or occurrence of an event in its environment?" The ghosts are certainly approaching Pac-man so the answer is Yes. Therefore, we move to Level 1.1 and ask: "Does the system have different sets of goals active during different environmental or bodily conditions?" If Pac-man eats a power-up pellet, the ghosts will change from approaching him to running away, so the answer is again Yes. Therefore, we go to Level 2.0 and ask: "Does the system develop new adaptive approach or avoidance patterns over time?" If we observe the ghosts long enough, we will notice that they essentially only have two behaviors and these are inflexible, predictable, and unchanging. Therefore the answer to this question is negative, which means that a Pac-man ghost is only a Level 1.1 system: a modal value-driven automaton.

S2.2 Example 2: A Chimpanzee

We now repeat the evaluation process for an entity with, arguably, a much higher degree of volition: a chimpanzee. Beginning with an (admittedly absurd) assumption of Level 0.0, we move to Level 0.1 and ask: "Does the system move or act on its own, i.e., without continuous prompting by external forces?" Anyone who's been to a zoo could answer Yes, so we move to Level 0.2 and ask: "Is the system's spontaneous behavior modified by events/conditions in the environment?" Watching chimpanzees playing in their enclosure quickly leads to a Yes conclusion, so we move to Level 1.0 and ask: "Does the system appear to be trying to approach or avoid any object or occurrence of an event in its

environment?" Chimps will approach food or other chimps and will sometimes avoid each other, so the answer is Yes. Therefore, we go to Level 1.1 and ask: "Does the system have different sets of goals active during different environmental or bodily conditions?" As chimps may forage, mate, fight, or do any number of complex behaviors that depend on their internal state and their environment, the answer is clearly Yes, so we move to Level 2.0 and ask: "Does the system develop new adaptive approach or avoidance patterns over time?" As shown by Wolfgang Köhler in the 1910s, chimps can learn to adopt such strategies as piling boxes under a banana tree and climbing on them to reach the fruit (Matsuzawa, 2002), so the answer is Yes. We now move to Level 2.1 and ask: "Can the system engage in a task that requires working memory (e.g. delayed non-match-to-sample)?" Chimps are capable of performing nearly perfectly on a visual delayed-match-to-sample task (Hashiya & Kojima, 2001) so the answer is Yes. We now move to Level 2.2 and ask: "Can the system engage in a task that requires long-term memory?" Chimps in the forests of Bossou forage for figs which means, they need to be able to remember the location of fig trees, the time of year that the fruits are ripe, the fact that red fruit is ripe whereas green fruit is unripe, and the best climbing routes to reach fruit in the highest trees (Matsuzawa, 2002). These memories would seem to require some form of long-term storage, so we answer Yes. Now we move to Level 2.3 and ask: "Can the system engage in a behavior (e.g. game-playing, navigation) that requires evaluation of multiple possibilities without action?" The behavioral pause that Köhler observed before his chimps engaged in the box-stacking, fruit-reaching behaviors, suggests that some kind of speculative deliberation was happening. Also, it was observed in 1970 that a chimp (named Julia) was able to choose the first time the correct key for opening boxes with other keys that opened other boxes finally leading to a box being opened with a food reward (Suddendorf & Corballis, 1997). This suggests that chimps can engage in the kind of Popperian reasoning we associate with deliberation, at least in a short time-frame, so we answer Yes. Now, we can move to Level 3.0 and ask: "Does the organism send and selectively respond to social cues?" The answer is an obvious Yes, so we go to Level 3.1 and ask: "Can the system pick up and move around objects in its environment?" Under the right conditions, the answer is observably Yes, such as when chimps use sticks to forage for termites (Dennett, 1996). Now,

we move to Level 3.2 and ask: “Does the system communicate using language that has syntax as well as semantics?” Although chimps such as Kanzi may be trained to use simple symbols structures, in the wild the answer generally appears to be No (Deacon, 1997), so we say that a chimp is a Level 3.1 system, i.e. a manipulative organism, though it’s not clear that chimps couldn’t be trained to achieve Level 3.2.

S3 Some Objections and Issues

There are a number of potential objections and issues to be addressed regarding the proposed scale and the methodology for using it. Some of the most salient of these follow.

S3.1 Volition and Consciousness

Usually, when we think of something being a volitional act, it means the act was voluntary. *That means that we consciously willed it. But many of these levels of so-called volition could be instantiated by “zombie” systems that aren’t aware of anything*. Little has been said thus far about consciousness and its relationship to volition. Degree of awareness was not built into either the definition of volition or the scale that has been proposed to measure it. One reason is that awareness is notoriously difficult to measure, except perhaps in humans who can be verbally instructed to self-report in natural language.

One of the most contentious arguments in philosophy of mind and AI has been over whether artificial systems could ever possess consciousness. Some philosophers such as John Searle are skeptical (Searle, 1980), while others (e.g. Churchland, Clark, Dennett) are more hopeful. Assuming a functionalist view of consciousness, as this paper does, however, allows one to take for granted that consciousness is a viable engineering problem.

Assuming functionalism to be true, both volition and consciousness are emergent properties of physical systems, and, in the view of this paper, both are best explained by graded scales rather than hard-threshold definitions. The lowest and (to the skeptic) least satisfying levels of volition may be possessed by creatures that are not self-aware as humans are and may even lack experience of qualia (though this is more debatable). However, both volition and consciousness increase as the information processing systems in an organism grow more complex.

Psychologist Daniel Wegner discusses the likely relation between volition and consciousness in humans in *The Illusion of Conscious Will* (2002). His central thesis is that "conscious will", i.e., the awareness of acting that we think causes our acts actually *follows* our acting and decision processes which can be regarded as unconscious. This view is supported by evidence from studies of neuroscience (Libet, 1999) and of cases where a sense of having willed an act are dissociated from the actual behavior of the person (e.g. "spirit possession", hypnotism, alien-hand syndrome, phantom limb movement of an amputee's absent limb). Consciousness of will may be an after-the-fact interpretation of what we actually are set to do!

However, even if this is so, the interpretation of the act is probably more than an epiphenomenon, as some would claim. Libet's own interpretation of his results (1999) is that the awareness of the behavior that follows the EEG-measured readiness potential and precedes the actual motor act may allow the motor act to be vetoed before execution (an idea the neurologist Ramachandran has dubbed "free won't" (Dennett, 2003)). It seems likely that conscious awareness may also be used to train the organism's future acts by, for example, reinforcing behaviors according to the emotional valence evoked by their execution and its consequences ("pride" or "guilt"). One thing this suggests is that consciousness is probably an integral part of the higher-level volitional processes such as deliberation. Automatic, well-learned, well-adapted responses do not require anything more than "zombie" processing, but when things go wrong or an organism is in a novel situation, then consciousness becomes important (Minsky, 1986).

In conclusion, lower levels of volitional processing may not be accompanied by consciousness (as most people understand it), but for higher levels—particularly those above Level 2.0—phenomenal consciousness may be an integral part of the decision process. If we are able to design a deliberative organism, for example, it will probably have a level of consciousness that is commensurate with that of non-primate mammals. As things stand, this would be notable progress for AI.

S3.2 Panpsychism

Humans and apes certainly seem to have volition, and probably other mammals. It's not so clear for fish and amphibians, and it is doubtful for insects. Clearly, however, thermostats, and at the other extreme, corporations, don't have any independent will. They are not conscious. They aren't even organisms! The approach taken in this paper does force us to at least reexamine our skeptical intuitions regarding the possibility of mind in non-animals and organizations—an earlier draft of the paper proposed collectives/societies of organisms as Level 4 entities—much in the same way the idea of extended mind forces us to consider the possible extension of mind into the environment. Acceptance of a functionalist position on materialism has the corollary that if inanimate matter or collections of individuals are organized in the right fashion, there should be consciousness, volition, and other mental attributes that accrue. This could be said to be a kind of "soft panpsychist" view of the universe, i.e., not everything in the universe necessarily possesses a mind, but (physical) systems that are sufficiently complex and organized so as to have an information processing structure analogous with that underlying animal consciousness, must be considered aware. We do not know what it's like to be a thermostat, or even (surprisingly, perhaps) a corporation or a crowd because each of our awarenesses is attached to a particular network of neurons and associated body. Even anything we know about the awareness of other human minds is inferred. Yet, it may be epistemologically useful to take the intentional stance regarding corporations and even thermostats, much as we often do with genes and memes (Dawkins, 1976). It may even be possible that, counter to our intuitions, that thermostats and societies *really are* conscious, willful, etc., in that they are at least dimly aware and qualia-experiencing beings. As we come to better understand the architecture of mind, our intuitions may (or may not) change regarding what kinds of entities have minds.

S3.3 Linear, Ordinal Scale Issues

Something seems rather arbitrary about a unidimensional linear scale for volition. It seems like it would be better to have multiple dimensions and a measure for each of these. This may be one of the most legitimate objections to the adoption of such a scale as proposed in this paper. However, if we are

willing to give up strong ontological claims about the particular analysis that has been chosen in this paper for volition, the scale may still be a valuable epistemological and methodological tool. (A Dennettian move of saying that the scale formalisms are, ontologically speaking, at least "real patterns" seems feasible (Dennett, 1991).)

There are any number of ways that the complex phenomenon of volition could be analyzed, much as there are any number of ways one might analyze the physical organization of a bacterium, e.g. a molecular vs. a cellular vs. an evolutionary or teleological analysis. In a way similar to Dennett's *Kinds of Minds* (1996) scale, this scale attempts to roughly follow what seems to have been evolution's course in the animal kingdom from simple invertebrate organisms to symbol-using primates. At risk of anthropocentrism, the following hierarchy was assumed: insects, fish, amphibians, reptiles, non-social mammals (e.g. tigers), social, but non-manipulating mammals (e.g. horses, dogs), non-human primates, pre-literate man, and literate man. Future ethological findings may suggest reconsiderations of this assumed hierarchy, but it seems like a plausible foundation on which to build the proposed volitional hierarchy.

Even given this animal hierarchy, however, some simplifications were made for functional tidiness. Many of the volitional functional features probably evolved into existence together rather than consecutively in the ordering given. Reinforcement learning, as noted, evolved with the biological innovation of neurons. Short-term and long-term memory probably developed in parallel, as well as the attentional mechanisms that are needed to control them for deliberation. Social and manipulative scaffolding probably evolved in parallel. Many insects, of course, are both social, and organized in societies.

While all of this may suggest that a unidimensional scale for volition may not be the most accurate detailed conceptual model of the phenomenon of volition, such an approach has many epistemological and methodological advantages for the study of volitional behavior in animals and artificial systems. It provides a comprehensive list of architectural features needed for engineering of human volition; even if we disagree on the ordering given for inclusion of the features, the list is useful.

This list and its ordering suggest a plan of attack for the analysis and engineering of volitional systems. The breakdown of volition into components encourages researchers to pinpoint corresponding mechanisms in the brain and suggests to AI researchers an ordering in which to build layers of volition with simpler layers being developed and tested before the more complex layers. (For example, we might conclude that we should develop deliberative organisms before we attempt to build symbolic organisms, i.e., systems that are capable of true natural language understanding.) Finally, the scale may be a useful evaluative tool both for natural and for artificial systems. It may provide a set of empirically defined benchmarks for AI for evaluating its progress in developing intelligent systems. These benchmarks, it is hoped, connect in a more satisfying way both to formal-philosophical and to commonsense ideas of volition, than the Turing test connects to concepts of intelligence.

S3.4 Volition and Mechanistic Causality

The kind of machine (or natural) volition the scale proposes doesn't sound much like true *volition because there is a questionable underlying assumption that you can have truly volitional acts from systems governed entirely by mechanistic principles*. This is an expected argument from (some) philosophers who take an *incompatibilist* stance, i.e., who don't believe free will is compatible with determinism. The intuition is that if a behavior is deterministically caused, it can't be free, and, therefore wasn't *truly* willed. Even an indeterminist, though, might ask whether randomness will make a system's choices any freer in a sense we would care about.

For Kane (1996), who is an incompatibilist *libertarian*—i.e., someone who thinks determinism is false and free will exists—the critical issue is whether the agent has ultimate responsibility for at least some of its acts, and he takes the position that indeterministic causation within the choice processes of the organism would allow this, whereas a deterministic process would not. Dennett (2003) disagrees and argues that deterministic unpredictability in the decision processes would give the same results in terms of both behavior and subjective experience.

Whether determinism or indeterminism holds, however, it makes sense to say that the acts of agents are caused by *something* and agents are not "prime movers unmoved". To say that volitional

behavior can't be explained by causal mechanisms is like saying a computer's behavior can't be explained by the activities of a lot of integrated circuits that are wired together. The main difference between the arguments is that we *know* how interconnected chips give rise to computer functionality because we designed computers, but in the case of humans, natural process has engineered them and we are thrust into the position of trying to reverse-engineer. If our knowledge of computers were forgotten and it were viewed as unethical to open up computers and see how they work or analyze source-code for software, we might imagine an observer of a PC running Windows 98 attributing the behavior of the machine to any number of mysterious acausal forces!

One key argument of compatibilists is that neither acausality nor *absolute, uninfluenced* agent responsibility are necessary for volitional exercise of the will. For those unwilling to concede this point, then the proposed scale might be viewed as a functional specification for causal systems that *simulate* the behaviors of systems that have *true* volition. But if the systems built according to these principles behave enough like real organisms, it will call into question the utility of making a distinction between simulated and real volition.

S3.5 Non-specification of Mechanisms

An abstract functional scale like this is useful to have as a benchmark maybe, but it doesn't tell us enough about specific mechanisms or algorithms. AI researchers are primarily interested in mechanisms, so indeed the scale is no detailed blueprint for a volitional organism. However, it does provide an "aerial view" of the specifications, and even suggests an order in which investigations might be made into mechanisms and algorithms. It may provide a framework, also, for interpreting the findings in ethology and neuroscience. Figuring out which areas of the human brain are implicated in the different levels of volition and then studying the mechanisms of those areas, may allow neurally-inspired AI researchers to construct more detailed specifications for the implementation of volitional systems.

Endnotes

1. Forward-chain reasoning may be a more common solution, but this can be accommodated within the same attention-directed short- and long-term memory system as the backward-chaining example.

References

Baddeley, A. D. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience, 4*, 829-839.

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

Cammaerts, M. C. (2004). Operant conditioning in the ant Myrmica sabuleti. *Behavioural Processes, 67*, 417-425.

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

Dautenhahn, K. (2000). Evolvability, Culture, and the Primate Brain. In C. L. Nehaniv (Ed.), *Proceedings of the Evolvability Workshop at the Seventh International Conference on the Simulation and Synthesis of Living Systems (Artificial Life VII), August 2000* (pp. 23-26).

Dawkins, R. (1976). *The Selfish Gene*.New York, NY: Oxford University Press.

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. (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.

Hashiya, K., & Kojima, S. (2001). Acquisition of auditory-visual intermodal matching-to-sample by a chimpanzee (Pan troglodytes): comparison with visual-visual intramodal matching. *Animal Cognition, 4*, 231-239.

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

Libet, B. (1999). Do we have free will? *Journal of Consciousness Studies, 6*(8-9), 47-57.

Mataric, M. (1991). Navigating with a rat brain: a neurobiologically inspired model for robot spatial representation. In J. A. Meyer & S. Wilson (Eds.), *From Animals to Animats I* (pp. 169-175). Cambridge, MA: MIT Press.

Matsuzawa, T. (2002). Chimpanzee Ai and her son Ayumu: an episode of education by master-apprenticeship. In M. Bekoff, C. Allen & G. M. Burhgardt (Eds.), *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition* (pp. 189-195). Cambridge, MA: MIT Press.

McDermott, D. (1997). How Intelligent is Deep Blue? Retrieved January 12, 2008, from http://www.psych.utoronto.ca/users/reingold/courses/ai/cache/mcdermott.html

Minsky, M. (1986). *The Society of Mind*.New York, NY: Simon and Schuster.

Papaj, D. R., & Lewis, A. C. (Eds.). (1992). *Insect Learning: Ecological and Evolutionary Perspectives*.Boston, MA: Kluwer Academic Publishers.

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

Searle, J. R. (1980). Minds, brains, and programs. In J. Haugeland (Ed.), *Mind Design II* (pp. 183-204). Cambridge, MA: MIT Press.

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.

Suddendorf, T., & Corballis, M. C. (1997). Mental time travel and the evolution of the human mind. *Genetic, Social, and General Psychology Monographs, 123*(2), 133-167.

Suri, R. E. (2002). TD models of reward predictive responses in dopamine neurons. *Neural Networks, 15*, 523-533.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*.Cambridge, MA: MIT Press.

Wegner, D. (2002). *The Illusion of Conscious Will*.Cambridge, MA: MIT Press.

Wersing, A. C., Grünewald, B., & Menzel, R. (2003). Cellular mechanisms of odor learning in honeybees. *Society for Neuroscience Abstract*.